

High-dimensional quadratic classifiers in non-sparse settings

Makoto Aoshima and Kazuyoshi Yata

Institute of Mathematics, University of Tsukuba, Ibaraki, Japan

Abstract

We consider high-dimensional quadratic classifiers in non-sparse settings. The target of classification rules is not Bayes error rates in the context. The classifier based on the Mahalanobis distance does not always give a preferable performance even if the populations are normal distributions having known covariance matrices. The quadratic classifiers proposed in this paper draw information about heterogeneity effectively through both the differences of expanding mean vectors and covariance matrices. We show that they hold a consistency property in which misclassification rates tend to zero as the dimension goes to infinity under non-sparse settings. We verify that they are asymptotically distributed as a normal distribution under certain conditions. We also propose a quadratic classifier after feature selection by using both the differences of mean vectors and covariance matrices. Finally, we discuss performances of the classifiers in actual data analyses. The proposed classifiers achieve highly accurate classification with very low computational costs.

Keywords: Bayes error rate; Discriminant analysis; Feature selection; Heterogeneity; Large p small n

1 Introduction

Globally, there is an ever increasing need for fast, accurate and cost effective analysis of high-dimensional data in many fields, including academia, medicine and business. However, existing classifiers for high-dimensional data are often complex, time consuming and have no guarantee of accuracy. In this paper we hope to provide better options. A common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively low. This is the so-called “HDLSS” or “large p , small n ” data situation where $p/n \rightarrow \infty$; here p is the data dimension and n is the sample size. Suppose we have independent and p -variate two populations, π_i , $i = 1, 2$, having an unknown mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{ip})^T$ and unknown covariance matrix $\boldsymbol{\Sigma}_i (> \mathbf{O})$ for each i . Let

$$\boldsymbol{\mu}_{12} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (\mu_{121}, \dots, \mu_{12p})^T \quad \text{and} \quad \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2.$$

We assume that $\limsup_{p \rightarrow \infty} |\mu_{12j}| < \infty$ for all j . Note that $\limsup_{p \rightarrow \infty} \|\boldsymbol{\mu}_{12}\|^2/p < \infty$, where $\|\cdot\|$ denotes the Euclidean norm. Let $\sigma_{i(j)}$ be the j -th diagonal element of $\boldsymbol{\Sigma}_i$ for $j = 1, \dots, p$ ($i = 1, 2$). We assume that $\sigma_{i(j)} \in (0, \infty)$ as $p \rightarrow \infty$ for all i, j . Here, for a function, $f(\cdot)$, “ $f(p) \in (0, \infty)$ as $p \rightarrow \infty$ ” implies that $\liminf_{p \rightarrow \infty} f(p) > 0$ and $\limsup_{p \rightarrow \infty} f(p) < \infty$. Then, it holds that $\text{tr}(\boldsymbol{\Sigma}_i)/p \in (0, \infty)$ as $p \rightarrow \infty$ for $i = 1, 2$. We

Address correspondence to Makoto Aoshima, Institute of Mathematics, University of Tsukuba, Ibaraki 305-8571, Japan; Fax: +81-298-53-6501; E-mail: aoshima@math.tsukuba.ac.jp

do not assume $\Sigma_1 = \Sigma_2$. The eigen-decomposition of Σ_i is given by $\Sigma_i = \mathbf{H}_i \mathbf{\Lambda}_i \mathbf{H}_i^T$, where $\mathbf{\Lambda}_i = \text{diag}(\lambda_{i1}, \dots, \lambda_{ip})$ is a diagonal matrix of eigenvalues, $\lambda_{i1} \geq \dots \geq \lambda_{ip} > 0$, and $\mathbf{H}_i = [\mathbf{h}_{i1}, \dots, \mathbf{h}_{ip}]$ is an orthogonal matrix of the corresponding eigenvectors. We have independent and identically distributed (i.i.d.) observations, $\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}$, from each π_i , where $\mathbf{x}_{ik} = (x_{i1k}, \dots, x_{ipk})^T$, $k = 1, \dots, n_i$. We assume $n_i \geq 2$, $i = 1, 2$. Let $n_{\min} = \min\{n_1, n_2\}$. We estimate $\boldsymbol{\mu}_i$ and Σ_i by $\bar{\mathbf{x}}_{in_i} = (\bar{x}_{i1n_i}, \dots, \bar{x}_{ipn_i})^T = \sum_{k=1}^{n_i} \mathbf{x}_{ik}/n_i$ and $\mathbf{S}_{in_i} = \sum_{k=1}^{n_i} (\mathbf{x}_{ik} - \bar{\mathbf{x}}_{in_i})(\mathbf{x}_{ik} - \bar{\mathbf{x}}_{in_i})^T / (n_i - 1)$. Let $s_{in_i(j)}$ be the j -th diagonal element of \mathbf{S}_{in_i} for $j = 1, \dots, p$ ($i = 1, 2$).

In this paper, we consider high-dimensional quadratic classifiers in non-sparse settings. Let $\mathbf{x}_0 = (x_{01}, \dots, x_{0p})^T$ be an observation vector of an individual belonging to one of the two populations. Let $|\mathbf{M}|$ be the determinant of a square matrix \mathbf{M} . When π_i s are Gaussian, a Bayes optimal rule is given as follows: One classifies the individual into π_1 if

$$(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T \Sigma_1^{-1} (\mathbf{x}_0 - \boldsymbol{\mu}_1) - \log |\Sigma_2 \Sigma_1^{-1}| < (\mathbf{x}_0 - \boldsymbol{\mu}_2)^T \Sigma_2^{-1} (\mathbf{x}_0 - \boldsymbol{\mu}_2) \quad (1.1)$$

and into π_2 otherwise. Since $\boldsymbol{\mu}_i$ s and Σ_i s are unknown, one usually considers the following typical classifier:

$$(\mathbf{x}_0 - \bar{\mathbf{x}}_{1n_1})^T \mathbf{S}_{1n_1}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_{1n_1}) - \log |\mathbf{S}_{2n_2} \mathbf{S}_{1n_1}^{-1}| < (\mathbf{x}_0 - \bar{\mathbf{x}}_{2n_2})^T \mathbf{S}_{2n_2}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_{2n_2}).$$

The classifier usually converges to the Bayes optimal classifier when $n_{\min} \rightarrow \infty$ while p is fixed or $n_{\min}/p \rightarrow \infty$. However, in the HDLSS context, the inverse matrix of \mathbf{S}_{in_i} does not exist. When $\Sigma_1 = \Sigma_2$, Bickel and Levina (2004) considered an inverse matrix defined by only diagonal elements of the pooled sample covariance matrix. Fan and Fan (2008) considered a classification after feature selection. Fan, Feng and Tong (2012) proposed the regularized optimal affine discriminant (ROAD). When $\Sigma_1 \neq \Sigma_2$, Dudoit, Fridlyand and Speed (2002) considered an inverse matrix defined by only diagonal elements of \mathbf{S}_{in_i} . Aoshima and Yata (2011) considered using $\{\text{tr}(\mathbf{S}_{in_i})/p\}^{-1} \mathbf{I}_p$ instead of $\mathbf{S}_{in_i}^{-1}$ from a geometrical background of HDLSS data and proposed the geometric classifier. Here, \mathbf{I}_p denotes the identity matrix of dimension p . Hall, Marron and Neeman (2005) and Marron, Todd and Ahn (2007) considered distance weighted classifiers. Chan and Hall (2009) and Aoshima and Yata (2014) considered distance-based classifiers and Aoshima and Yata (2014) gave the misclassification rate adjusted classifier for multiclass, high-dimensional data whose misclassification rates are no more than specified thresholds.

Recently, Cai and Liu (2011), Shao et al. (2011) and Li and Shao (2015) gave sparse linear or quadratic classification rules for high-dimensional data. They showed that their classification rules have Bayes error rates when π_i s are Gaussian. They assumed that λ_{ij} s are bounded under some sparsity conditions such as $\boldsymbol{\mu}_{12}$, Σ_i s and Σ_{12} (or Σ_i^{-1} s and $\Sigma_1^{-1} - \Sigma_2^{-1}$) are sparse. For example, when $\Sigma_1 = \Sigma_2 (= \Sigma, \text{ say})$, the error rate of their classification rules is given by $\Phi(-\Delta_{MD}^{1/2}/2) + o(1)$ as $p \rightarrow \infty$, where $\Delta_{MD} = \boldsymbol{\mu}_{12}^T \Sigma^{-1} \boldsymbol{\mu}_{12}$ that is the Mahalanobis distance and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Here, $\Phi(-\Delta_{MD}^{1/2}/2)$ is the Bayes error rate.

In this paper, we investigate quadratic classifiers from a perspective that is different from the sparse discriminant analysis. We *do not assume that $\boldsymbol{\mu}_{12}$, Σ_i s and Σ_{12} are sparse*. In such a context, the target of classification rules is not Bayes error rates as in $\Phi(-\Delta_{MD}^{1/2}/2) + o(1)$ as $p \rightarrow \infty$. We consider a consistency property such as misclassification

rates tend to 0 as p increases, i.e.,

$$e(i) \rightarrow 0 \text{ as } p \rightarrow \infty \text{ for } i = 1, 2,$$

where $e(i)$ denotes the error rate of misclassifying an individual from π_i into the other class. For example, if one can assume that π_i s are Gaussian and $\Sigma_1 = \Sigma_2$, the Bayes rule by (1.1) has such a consistency property when $\Delta_{MD} \rightarrow \infty$ as $p \rightarrow \infty$. It is likely that $\Delta_{MD} \rightarrow \infty$ as $p \rightarrow \infty$ when μ_{12} is non-sparse in the sense that $\|\mu_{12}\| \rightarrow \infty$ as $p \rightarrow \infty$. We emphasize that such non-sparse situations often occur in high-dimensional settings. For example, see Hall, Marron and Neeman (2005) or (6.1), (6.2) and Table 2 in Section 6. We will show that quadratic classifiers hold the consistency property when μ_{12} or Σ_{12} is non-sparse such as $\|\mu_{12}\| \rightarrow \infty$ or $\|\Sigma_{12}\|_F \rightarrow \infty$ as $p \rightarrow \infty$, where $\|\cdot\|_F$ is the Frobenius norm.

In this paper, we consider the following function of \mathbf{A}_i to discriminate π_i s in general:

$$W_i(\mathbf{A}_i) = (\mathbf{x}_0 - \bar{\mathbf{x}}_{in_i})^T \mathbf{A}_i (\mathbf{x}_0 - \bar{\mathbf{x}}_{in_i}) - \text{tr}(\mathbf{S}_{in_i} \mathbf{A}_i) / n_i - \log |\mathbf{A}_i|, \quad (1.2)$$

where \mathbf{A}_i is a positive definite matrix satisfying the equation that $\text{tr}\{\Sigma_i(\mathbf{A}_{i'} - \mathbf{A}_i)\} = \text{tr}(\mathbf{A}_i^{-1} \mathbf{A}_{i'}) - p$ ($i \neq i'$). Here, $\text{tr}(\mathbf{S}_{in_i} \mathbf{A}_i) / n_i$ is a bias correction term. We consider a quadratic classification rule in which one classifies the individual into π_1 if

$$W_1(\mathbf{A}_1) - W_2(\mathbf{A}_2) < 0 \quad (1.3)$$

and into π_2 otherwise. Note that (1.3) becomes a linear classifier when $\mathbf{A}_1 = \mathbf{A}_2$. We have that $E\{W_{i'}(\mathbf{A}_{i'})\} - E\{W_i(\mathbf{A}_i)\} = \Delta_i$ when $\mathbf{x}_0 \in \pi_i$, where

$$\Delta_i = \mu_{12}^T \mathbf{A}_{i'} \mu_{12} + \text{tr}\{\Sigma_i(\mathbf{A}_{i'} - \mathbf{A}_i)\} + \log |\mathbf{A}_{i'}^{-1} \mathbf{A}_i| \quad (1.4)$$

for $i = 1, 2$ ($i' \neq i$).

Proposition 1.1. (i) $\Delta_i \geq 0$. (ii) $\Delta_i > 0$ when $\mu_1 \neq \mu_2$ or $\mathbf{A}_1 \neq \mathbf{A}_2$.

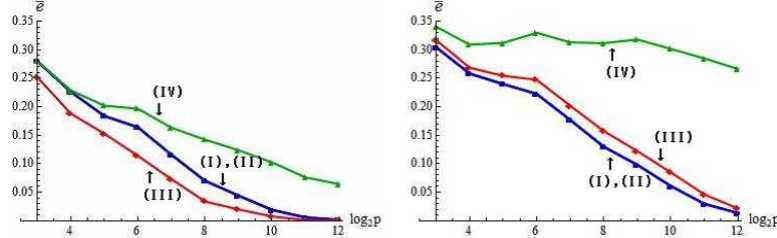
Remark 1. As for l (≥ 3)-class classification, one may consider a classification rule such as one classifies the individual into π_i if

$$\underset{i'=1, \dots, l}{\text{argmin}} W_{i'}(\mathbf{A}_{i'}) = i.$$

In this paper, we specially consider the following four typical \mathbf{A}_i s in (1.2):

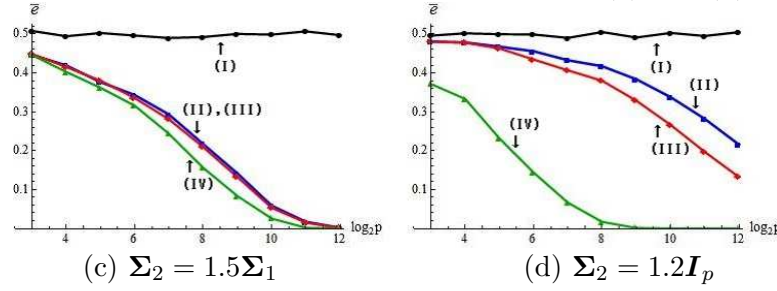
$$\text{(I) } \mathbf{A}_i = \mathbf{I}_p, \quad \text{(II) } \mathbf{A}_i = \frac{p}{\text{tr}(\Sigma_i)} \mathbf{I}_p, \quad \text{(III) } \mathbf{A}_i = \Sigma_{i(d)}^{-1}, \quad \text{and} \quad \text{(IV) } \mathbf{A}_i = \Sigma_i^{-1},$$

where $\Sigma_{i(d)} = \text{diag}(\sigma_{i(1)}, \dots, \sigma_{i(p)})$. These four \mathbf{A}_i s satisfy the condition that $\text{tr}\{\Sigma_i(\mathbf{A}_{i'} - \mathbf{A}_i)\} = \text{tr}(\mathbf{A}_i^{-1} \mathbf{A}_{i'}) - p$ ($i \neq i'$) and they provide historical background of discriminant analysis. Note that $\|\Sigma_{12}\|_F \geq \|\mathbf{A}_1^{-1} - \mathbf{A}_2^{-1}\|_F$ for these four \mathbf{A}_i s. Also, under (I) to (IV), we note that $\Delta_i \rightarrow \infty$ as $p \rightarrow \infty$ when μ_{12} or Σ_{12} is non-sparse. Practically, \mathbf{A}_i s should be estimated except for (I). We will consider quadratic classifiers given by estimating \mathbf{A}_i s in Section 4. Now, let us see an easy example to check the performance of (I) to (IV) in (1.3). We set $p = 2^s$, $s = 3, \dots, 12$. Independent pseudo random observations were generated from $\pi_i : N_p(\mu_i, \Sigma_i)$, $i = 1, 2$. We set $\mu_1 = \mathbf{0}$ and $\Sigma_1 = \mathbf{B}_1(0.3^{|i-j|^{1/3}})\mathbf{B}_1$, where $\mathbf{B}_1 = \text{diag}\{[0.5 + 1/(p+1)]^{1/2}, \dots, [0.5 + p/(p+1)]^{1/2}\}$. Note that $\text{tr}(\Sigma_1) = p$ and $\Sigma_{1(d)} = \mathbf{B}_1^2$. When $\Sigma_1 = \Sigma_2$ and $(n_1, n_2) = (\log_2 p, 2 \log_2 p)$, we considered two cases:



(a) $\mu_1 = \mathbf{0}$ and $\mu_2 = (1, \dots, 1, 0, \dots, 0)^T$ (b) $\mu_1 = \mathbf{0}$ and $\mu_2 = (0, \dots, 0, 1, \dots, 1)^T$

Figure 1: The average error rates of the classification rule by (1.3) for (I) to (IV) when $\Sigma_1 = \Sigma_2$. The left and right panels display \bar{e} in the cases of (a) and (b), respectively.



(c) $\Sigma_2 = 1.5\Sigma_1$

(d) $\Sigma_2 = 1.2I_p$

Figure 2: The average error rates of the classification rule by (1.3) for (I) to (IV) when $\mu_1 = \mu_2$. The left and right panels display \bar{e} in the cases of (c) and (d), respectively.

- (a) $\mu_2 = (1, \dots, 1, 0, \dots, 0)^T$ whose first $\lceil p^{2/3} \rceil$ elements are 1, and
- (b) $\mu_2 = (0, \dots, 0, 1, \dots, 1)^T$ whose last $\lceil p^{2/3} \rceil$ elements are 1.

Here, $\lceil x \rceil$ denotes the smallest integer $\geq x$. Next, when $\mu_2 = \mathbf{0}$ (i.e., $\mu_{12} = \mathbf{0}$) and $(n_1, n_2) = (5, 10)$, we considered two cases:

- (c) $\Sigma_2 = 1.5\Sigma_1$ and (d) $\Sigma_2 = 1.2I_p$.

Note that μ_{12} or Σ_{12} is non-sparse for (a) to (d) because $\|\mu_{12}\| \rightarrow \infty$ or $\|\Sigma_{12}\|_F \rightarrow \infty$ as $p \rightarrow \infty$. For $\mathbf{x}_0 \in \pi_i$ ($i = 1, 2$) we repeated 2000 times to confirm if the classification rule by (1.3) with either of (I) to (IV) does (or does not) classify \mathbf{x}_0 correctly and defined $P_{ir} = 0$ (or 1) accordingly for each π_i . We calculated the error rates, $\bar{e}(i) = \sum_{r=1}^{2000} P_{ir}/2000$, $i = 1, 2$. Also, we calculated the average error rate, $\bar{e} = \{\bar{e}(1) + \bar{e}(2)\}/2$. Their standard deviations are less than 0.011. In Figure 1, we plotted \bar{e} for (a) and (b). Note that (I) is equivalent to (II) for (a) and (b). In Figure 2, we plotted \bar{e} for (c) and (d). We observed that (IV) gives the worst performance in Figure 1 contrary to expectations. In general, one would think that the classifier based on the Mahalanobis distance such as (1.2) with (IV) is the best when π_i s are Gaussian and $n_{\min} \rightarrow \infty$. We emphasize that it is not true for high-dimensional data. We will explain its theoretical reason in Section 3.2. We observed that (I) (or (II)) gives a better performance compared to (III) for (b) in Figure 1. We will discuss the reasons in Section 3.4. In Figure 2, the error rates of (I) are close to 0.5 because of $\mu_{12} = \mathbf{0}$. On the other hand, (II), (III) and (IV) gave good performances as p increases by drawing information on heteroscedasticity in the classifiers. We will give their theoretical backgrounds in Sections 2.2 and 3.4.

We pay special attention to the difference of covariance matrices in classification for high-dimensional data. In Section 2, we show that the classification rule by (1.3) holds

the consistency property under non-sparse settings. In Section 3, we verify that the quadratic classifier by (1.2) is asymptotically distributed as a normal distribution under certain conditions. In Section 4, we consider the estimation of \mathbf{A}_i s and give asymptotic properties of estimated classifiers. In Section 5, we propose a quadratic classifier after feature selection by using both the differences of mean vectors and covariance matrices. In Section 6, we discuss performances of the classifiers in actual data analyses. Finally, in Section 7, we give concluding remarks of our study.

2 Consistency property of the quadratic classifier

In this section, we discuss the consistency property of quadratic classifiers given by (1.2).

2.1 Preliminary

Similar to Bai and Saranadasa (1996) and Aoshima and Yata (2014), we assume the following assumption about population distributions as necessary:

(A-i) Let \mathbf{y}_{ik} , $k = 1, \dots, n_i$, be i.i.d. random q_i -vectors having $E(\mathbf{y}_{ik}) = \mathbf{0}$ and $\text{Var}(\mathbf{y}_{ik}) = \mathbf{I}_{q_i}$ for each i ($= 1, 2$), where $q_i \geq p$. Let $\mathbf{y}_{ik} = (y_{i1k}, \dots, y_{iq_ik})^T$ whose components satisfy that $\limsup_{p \rightarrow \infty} E(y_{ijk}^4) < \infty$ for all j and

$$E(y_{ijk}^2 y_{irk}^2) = E(y_{ijk}^2) E(y_{irk}^2) = 1 \quad \text{and} \quad E(y_{ijk} y_{irk} y_{isk} y_{itk}) = 0 \quad (2.1)$$

for all $j \neq r, s, t$. Then, the observations, \mathbf{x}_{ik} s, from each π_i ($i = 1, 2$) are given by

$$\mathbf{x}_{ik} = \mathbf{\Gamma}_i \mathbf{y}_{ik} + \boldsymbol{\mu}_i, \quad k = 1, \dots, n_i,$$

where $\mathbf{\Gamma}_i = [\boldsymbol{\gamma}_{i1}, \dots, \boldsymbol{\gamma}_{iq_i}]$ is a $p \times q_i$ matrix such that $\mathbf{\Gamma}_i \mathbf{\Gamma}_i^T = \boldsymbol{\Sigma}_i$.

Note that $\mathbf{\Gamma}_i$ includes the case that $\mathbf{\Gamma}_i = \mathbf{H}_i \boldsymbol{\Lambda}_i^{1/2} = [\lambda_{i1}^{1/2} \mathbf{h}_{i1}, \dots, \lambda_{ip}^{1/2} \mathbf{h}_{ip}]$. We assume the following assumption instead of (A-i) as necessary:

(A-ii) (A-i) by replacing (2.1) with the independence of y_{ijk} , $j = 1, \dots, q_i$ ($i = 1, 2$; $k = 1, \dots, n_i$).

Note that (A-ii) is a special case of (A-i). When π_i has $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, (A-ii) naturally holds.

Now, we consider the following divergence condition for p and n_i s:

(\star) $p \rightarrow \infty$ either when n_i is fixed or $n_i \rightarrow \infty$ for $i = 1, 2$.

Let $\Delta_{iA} = \boldsymbol{\mu}_{12}^T \mathbf{A}_{i'} \boldsymbol{\Sigma}_i \mathbf{A}_i \boldsymbol{\mu}_{12}$ for $i = 1, 2$ ($i' \neq i$). We consider the following conditions under (\star) for $i = 1, 2$ ($i' \neq i$):

$$\text{(C-i)} \quad \frac{\text{tr}\{(\boldsymbol{\Sigma}_i \mathbf{A}_i)^2\}}{n_i \Delta_i^2} = o(1) \quad \text{and} \quad \frac{\text{tr}(\boldsymbol{\Sigma}_i \mathbf{A}_{i'} \boldsymbol{\Sigma}_{i'} \mathbf{A}_{i'}) + \text{tr}\{(\boldsymbol{\Sigma}_{i'} \mathbf{A}_{i'})^2\}/n_{i'}}{n_{i'} \Delta_i^2} = o(1),$$

$$\text{(C-ii)} \quad \frac{\Delta_{iA}}{\Delta_i^2} = o(1), \quad \text{and} \quad \text{(C-iii)} \quad \frac{\text{tr}[\{\boldsymbol{\Sigma}_i(\mathbf{A}_1 - \mathbf{A}_2)\}^2]}{\Delta_i^2} = o(1).$$

Then, we claim the consistency property of (1.2) in (1.3) as follows:

Theorem 2.1. Assume (A-i). Assume also (C-i) to (C-iii). Then, we have that

$$\frac{W_{i'}(\mathbf{A}_{i'}) - W_i(\mathbf{A}_i)}{\Delta_i} = 1 + o_P(1) \quad \text{under } (\star) \text{ when } \mathbf{x}_0 \in \pi_i \text{ for } i = 1, 2 \text{ (} i' \neq i \text{)}.$$

Furthermore, for the classification rule by (1.3) with (1.2), we have that

$$e(i) \rightarrow 0, \quad i = 1, 2, \quad \text{under } (\star). \quad (2.2)$$

Remark 2. When $\mathbf{A}_1 = \mathbf{A}_2$, we can claim Theorem 2.1 without (A-i) and (C-iii).

Let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ be the smallest and the largest eigenvalues of any positive definite matrix, \mathbf{M} . We use the phrase “ $\lambda(\mathbf{M}) \in (0, \infty)$ as $p \rightarrow \infty$ ” in the sense that $\liminf_{p \rightarrow \infty} \lambda_{\min}(\mathbf{M}) > 0$ and $\limsup_{p \rightarrow \infty} \lambda_{\max}(\mathbf{M}) < \infty$. We note that \mathbf{A}_i s in (I) to (III) satisfy the condition “ $\lambda(\mathbf{A}_i) \in (0, \infty)$ as $p \rightarrow \infty$ ”. Let $\Delta_{\min} = \min\{\Delta_1, \Delta_2\}$, $\lambda_{\max} = \max\{\lambda_{\max}(\Sigma_1), \lambda_{\max}(\Sigma_2)\}$ and $\text{tr}(\Sigma_{\max}^2) = \max\{\text{tr}(\Sigma_1^2), \text{tr}(\Sigma_2^2)\}$. Now, instead of (C-i) and (C-ii), we consider the following simpler conditions under (\star) :

$$(\mathbf{C-i}') \quad \frac{\text{tr}(\Sigma_{\max}^2)}{n_{\min} \Delta_{\min}^2} = o(1) \quad \text{and} \quad (\mathbf{C-ii}') \quad \frac{\lambda_{\max}}{\Delta_{\min}} = o(1).$$

Proposition 2.1. Assume that $\limsup_{p \rightarrow \infty} \lambda_{\max}(\mathbf{A}_i) < \infty$ for $i = 1, 2$. Then, (C-i') and (C-ii') imply (C-i) and (C-ii), respectively. Furthermore, if $\lambda(\mathbf{A}_i) \in (0, \infty)$ as $p \rightarrow \infty$ for $i = 1, 2$, and \mathbf{A}_i , $i = 1, 2$, are diagonal matrices such as in (I) to (III) in Section 1, (C-ii') implies (C-iii).

From the fact that $\lambda_{\max}(\Sigma_i) \leq \text{tr}(\Sigma_i^2)^{1/2}$ for $i = 1, 2$, we note that (C-i') and (C-ii') hold even when n_{\min} is fixed under

$$\text{tr}(\Sigma_{\max}^2)/\Delta_{\min}^2 \rightarrow 0 \quad \text{as } p \rightarrow \infty. \quad (2.3)$$

2.2 Consistency property for (I) to (IV)

As mentioned in Section 1, four typical \mathbf{A}_i s were specifically selected. For (I), by putting $\mathbf{A}_i = \mathbf{I}_p$, $i = 1, 2$, (1.2) and (1.4) are given as

$$\begin{aligned} W_i(\mathbf{I}_p) &= \|\mathbf{x}_0 - \bar{\mathbf{x}}_{in_i}\|^2 - \text{tr}(\mathbf{S}_{in_i})/n_i \\ \text{and } \Delta_1 &= \Delta_2 = \|\boldsymbol{\mu}_{12}\|^2 \quad (\text{hereafter called } \Delta_{(I)}). \end{aligned} \quad (2.4)$$

For (II), by putting $\mathbf{A}_i = \{p/\text{tr}(\Sigma_i)\}\mathbf{I}_p$, $i = 1, 2$, they are given as

$$\begin{aligned} W_i(\{p/\text{tr}(\Sigma_i)\}\mathbf{I}_p) &= \frac{p\|\mathbf{x}_0 - \bar{\mathbf{x}}_{in_i}\|^2}{\text{tr}(\Sigma_i)} - \frac{p\text{tr}(\mathbf{S}_{in_i})}{n_i\text{tr}(\Sigma_i)} + p \log\{\text{tr}(\Sigma_i)/p\} \\ \text{and } \Delta_i &= \frac{p\Delta_{(I)}}{\text{tr}(\Sigma_{i'})} + \frac{p\text{tr}(\Sigma_i)}{\text{tr}(\Sigma_{i'})} - p + p \log\left\{\frac{\text{tr}(\Sigma_{i'})}{\text{tr}(\Sigma_i)}\right\} \quad (\text{hereafter called } \Delta_{i(II)}). \end{aligned} \quad (2.5)$$

For (III), by putting $\mathbf{A}_i = \Sigma_{i(d)}^{-1}$, $i = 1, 2$, they are given as

$$\begin{aligned} W_i(\Sigma_{i(d)}^{-1}) &= \sum_{j=1}^p \left(\frac{(x_{0j} - \bar{x}_{ijn_i})^2}{\sigma_{i(j)}} - \frac{s_{in_i(j)}}{n_i\sigma_{i(j)}} + \log \sigma_{i(j)} \right) \\ \text{and } \Delta_i &= \sum_{j=1}^p \left\{ \frac{\mu_{12j}^2}{\sigma_{i'(j)}} + \frac{\sigma_{i(j)}}{\sigma_{i'(j)}} - 1 + \log\left(\frac{\sigma_{i'(j)}}{\sigma_{i(j)}}\right) \right\} \quad (\text{hereafter called } \Delta_{i(III)}). \end{aligned} \quad (2.6)$$

For (IV), by putting $\mathbf{A}_i = \mathbf{\Sigma}_i^{-1}$, $i = 1, 2$, they are given as

$$W_i(\mathbf{\Sigma}_i^{-1}) = (\mathbf{x}_0 - \bar{\mathbf{x}}_{in_i})^T \mathbf{\Sigma}_i^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_{in_i}) - \frac{\text{tr}(\mathbf{S}_{in_i} \mathbf{\Sigma}_i^{-1})}{n_i} + \sum_{j=1}^p \log \lambda_{ij} \quad (2.7)$$

$$\text{and } \Delta_i = \boldsymbol{\mu}_{12}^T \mathbf{\Sigma}_{i'}^{-1} \boldsymbol{\mu}_{12} + \text{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1}) - p + \sum_{j=1}^p \log \left(\frac{\lambda_{i'j}}{\lambda_{ij}} \right) \quad (\text{hereafter called } \Delta_{i(IV)}).$$

We first consider the classifiers by (2.4) to (2.6). From Theorem 2.1 and Proposition 2.1, we have the following result.

Corollary 2.1. *Assume (C-i') and (C-ii'). Then, for the classification rule by (1.3) with (2.4), we have (2.2). Furthermore, for the classification rule by (1.3) with (2.5) or (2.6), we have (2.2) under (A-i).*

We note that the classifier by (2.4) is equivalent to the distance-based classifier by Aoshima and Yata (2014). Hereafter, we call the classifier by (2.4) the “distance-based discriminant analysis (DBDA)”. From Corollary 2.1, under (2.3), the classification rule by (1.3) with (2.4), (2.5) or (2.6) has (2.2) even when n_i s are fixed. Note that DBDA has the consistency property without (A-i), so that DBDA is quite robust for non-Gaussian cases. See Aoshima and Yata (2014) for details. When $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, DBDA does not satisfy (C-i') and (C-ii'), on the other hand, the classifier by (2.5) or (2.6) still satisfies them.

Now, we consider the following condition for $\mathbf{\Sigma}_i$, $i = 1, 2$:

$$\text{tr}(\mathbf{\Sigma}_i^2)/\text{tr}(\mathbf{\Sigma}_i)^2 \rightarrow 0 \text{ as } p \rightarrow \infty. \quad (2.8)$$

We note that $\text{tr}(\mathbf{\Sigma}_i^2)/\text{tr}(\mathbf{\Sigma}_i)^2$ is a measure of sphericity. Also, note that (2.8) is equivalent to the condition that “ $\lambda_{\max}(\mathbf{\Sigma}_i)/\text{tr}(\mathbf{\Sigma}_i) \rightarrow 0$ as $p \rightarrow \infty$ ”. Under (A-i) and (2.8), from the fact that $\text{Var}(\|\mathbf{x}_0 - \boldsymbol{\mu}_i\|^2) = O\{\text{tr}(\mathbf{\Sigma}_i^2)\}$ when $\mathbf{x}_0 \in \pi_i$, we have that as $p \rightarrow \infty$

$$\|\mathbf{x}_0 - \boldsymbol{\mu}_i\| = \text{tr}(\mathbf{\Sigma}_i)^{1/2} \{1 + o_P(1)\} \text{ when } \mathbf{x}_0 \in \pi_i.$$

Thus the centroid data lies near the surface of an expanding sphere. See Hall, Marron and Neeman (2005) for details of the geometric representation. We emphasize that the classifier by (2.5) draws information about heteroscedasticity thorough the geometric representation having different radii, $\text{tr}(\mathbf{\Sigma}_i)^{1/2}$ s, of expanding two spheres. Note that $\text{tr}(\mathbf{\Sigma}_i^2) = o(p^2)$ under (2.8). Hence, for the classifier by (2.5), (2.3) holds under (2.8) and $\liminf_{p \rightarrow \infty} \Delta_{\min(II)}/p > 0$, where $\Delta_{\min(II)} = \min\{\Delta_{1(II)}, \Delta_{2(II)}\}$. Note that $\Delta_{\min(II)} > 0$ when $\text{tr}(\mathbf{\Sigma}_1) \neq \text{tr}(\mathbf{\Sigma}_2)$ in view of Proposition 1.1. If one can assume that $\liminf_{p \rightarrow \infty} |\text{tr}(\mathbf{\Sigma}_1)/\text{tr}(\mathbf{\Sigma}_2) - 1| > 0$, it follows $\liminf_{p \rightarrow \infty} \Delta_{\min(II)}/p > 0$, so that (2.3) holds under (2.8). Hence, for the classification rule by (1.3) with (2.5), we have (2.2) even when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and n_i s are fixed. See (II) in Figure 2. The accuracy becomes higher as the difference between $\text{tr}(\mathbf{\Sigma}_i)$ s grows.

Similarly, for the classifier by (2.6), it follows that (2.3) holds under (2.8) and $\liminf_{p \rightarrow \infty} \Delta_{\min(III)}/p > 0$, where $\Delta_{\min(III)} = \min\{\Delta_{1(III)}, \Delta_{2(III)}\}$. If one can assume that $\liminf_{p \rightarrow \infty} \sum_{j=1}^p |\sigma_{1(j)}/\sigma_{2(j)} - 1|/p > 0$, it follows $\liminf_{p \rightarrow \infty} \Delta_{\min(III)}/p > 0$, so that the classification rule by (1.3) with (2.6) has (2.2) even when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and n_i s are fixed. The classifier by (2.6) draws information about heteroscedasticity via the difference of diagonal elements

between the two covariance matrices. The accuracy becomes higher as the difference of those diagonal elements grows. See (III) in Figure 2.

Next, we consider the classifier by (2.7). From Theorem 2.1 and Proposition 2.1, we have the following result.

Corollary 2.2. *Assume (A-i). Assume also $\liminf_{p \rightarrow \infty} \lambda_{\min}(\mathbf{\Sigma}_i) > 0$ for $i = 1, 2$. Then, for the classification rule by (1.3) with (2.7), we have (2.2) under (C-i'), (C-ii') and the condition that $\text{tr}\{(\mathbf{I}_p - \mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1})^2\} = o(\Delta_{\min(IV)}^2)$ for $i = 1, 2$ ($i' \neq i$), where $\Delta_{\min(IV)} = \min\{\Delta_{1(IV)}, \Delta_{2(IV)}\}$.*

When $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, note that $\Delta_{\min(IV)} > 0$ in view of Proposition 1.1. Then, we have the following result.

Proposition 2.2. *When $\liminf_{p \rightarrow \infty} |\text{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1})/p - 1| > 0$ or $\liminf_{p \rightarrow \infty} \sum_{j=1}^p |\lambda_{ij}/\lambda_{i'j} - 1|/p > 0$ ($i \neq i'$), it follows that $\liminf_{p \rightarrow \infty} \Delta_{i(IV)}/p > 0$.*

Note that $\text{tr}\{(\mathbf{I}_p - \mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1})^2\} \leq p + \text{tr}\{(\mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1})^2\} = p + O\{\text{tr}(\mathbf{\Sigma}_i^2)\} = o(p^2)$ under (2.8) and $\liminf_{p \rightarrow \infty} \lambda_{\min}(\mathbf{\Sigma}_{i'}) > 0$. Hence, from Corollary 2.2, for the classification rule by (1.3) with (2.7), we have (2.2) under (A-i), (2.8), $\liminf_{p \rightarrow \infty} \Delta_{\min(IV)}/p > 0$ and $\liminf_{p \rightarrow \infty} \lambda_{\min}(\mathbf{\Sigma}_i) > 0$ for $i = 1, 2$. Thus from Proposition 2.2, the accuracy becomes higher as the difference of eigenvalues or eigenvectors between the two covariance matrices grows. See (IV) in Figure 2.

3 Asymptotic normality of the quadratic classifier

In this section, we discuss the asymptotic normality of quadratic classifiers given by (1.2). We further discuss Bayes error rates for high-dimensional data.

3.1 Preliminary

Let

$$\delta_i = 2 \left\{ \frac{\text{tr}\{(\mathbf{\Sigma}_i \mathbf{A}_i)^2\}}{n_i} + \frac{\text{tr}(\mathbf{\Sigma}_i \mathbf{A}_{i'} \mathbf{\Sigma}_{i'} \mathbf{A}_{i'})}{n_{i'}} + \Delta_{iA} \right\}^{1/2} \text{ for } i = 1, 2 \text{ } (i' \neq i).$$

Note that $\delta_i^2 = \text{Var}[2(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T \{\mathbf{A}_i(\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_i) - \mathbf{A}_{i'}(\bar{\mathbf{x}}_{i'n_{i'}} - \boldsymbol{\mu}_{i'} + (-1)^i \boldsymbol{\mu}_{12})\}]$ for $i = 1, 2$ ($i' \neq i$). Let $m = \min\{p, n_{\min}\}$. We assume the following conditions when $m \rightarrow \infty$ for $i = 1, 2$ ($i' \neq i$):

$$\text{(C-iv)} \quad \frac{\boldsymbol{\mu}_{12}^T \mathbf{A}_{i'} \mathbf{\Sigma}_{i'} \mathbf{A}_{i'} \boldsymbol{\mu}_{12} + \text{tr}\{(\mathbf{\Sigma}_{i'} \mathbf{A}_{i'})^2\}/n_{i'}}{n_{i'} \delta_i^2} = o(1), \quad \frac{\text{tr}\{(\mathbf{\Sigma}_i \mathbf{A}_i)^4\}}{n_i^2 \delta_i^4} = o(1) \text{ and}$$

$$\frac{\text{tr}\{(\mathbf{\Sigma}_i \mathbf{A}_{i'} \mathbf{\Sigma}_{i'} \mathbf{A}_{i'})^2\}}{n_{i'}^2 \delta_i^4} = o(1);$$

$$\text{(C-v)} \quad \frac{\text{tr}[\{\mathbf{\Sigma}_i(\mathbf{A}_1 - \mathbf{A}_2)\}^2]}{\delta_i^2} = o(1); \text{ and } \text{(C-vi)} \quad \frac{\Delta_{iA}}{\delta_i^2} = o(1).$$

From (A.6) in Appendix, under (A-i), (C-iv) and (C-v), it holds that

$$W_{i'}(\mathbf{A}_{i'}) - W_i(\mathbf{A}_i) - \Delta_i = 2(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T \left\{ \mathbf{A}_i(\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_i) - \mathbf{A}_{i'}(\bar{\mathbf{x}}_{i'n_{i'}} - \boldsymbol{\mu}_{i'} + (-1)^i \boldsymbol{\mu}_{12}) \right\} + o_P(\delta_i)$$

as $m \rightarrow \infty$ when $\mathbf{x}_0 \in \pi_i$ for $i = 1, 2$ ($i' \neq i$). Under (C-vi), it holds that $(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T \mathbf{A}_{i'} \boldsymbol{\mu}_{12} = o_P(\delta_i)$ as $m \rightarrow \infty$ when $\mathbf{x}_0 \in \pi_i$ for $i = 1, 2$ ($i' \neq i$). Then, we claim the asymptotic normality of (1.2) under (A-i) as follows:

Theorem 3.1. *Assume (A-i). Assume also (C-iv) to (C-vi). Then, we have that*

$$\frac{W_{i'}(\mathbf{A}_{i'}) - W_i(\mathbf{A}_i) - \Delta_i}{\delta_i} \Rightarrow N(0, 1) \quad \text{as } m \rightarrow \infty \quad (3.1)$$

when $\mathbf{x}_0 \in \pi_i$ for $i = 1, 2$ ($i' \neq i$),

where “ \Rightarrow ” denotes the convergence in distribution and $N(0, 1)$ denotes a random variable distributed as the standard normal distribution. Furthermore, for the classification rule by (1.3) with (1.2), it holds that

$$e(i) = \Phi\left(\frac{-\Delta_i}{\delta_i}\right) + o(1) \quad \text{as } m \rightarrow \infty \text{ for } i = 1, 2. \quad (3.2)$$

Let $\delta_{\min} = \min\{\delta_1, \delta_2\}$. Now, instead of (C-iv) to (C-vi), we consider the following conditions when $m \rightarrow \infty$:

$$(C\text{-iv}') \quad \frac{\|\boldsymbol{\mu}_{12}\|^2 \lambda_{\max} + \text{tr}(\boldsymbol{\Sigma}_{\max}^2)/n_{\min}}{n_{\min} \delta_{\min}^2} = o(1) \quad \text{and} \quad \frac{\lambda_{\max}^2}{n_{\min} \delta_{\min}^2} = o(1),$$

$$(C\text{-v}') \quad \frac{\text{tr}\{(\mathbf{A}_1 - \mathbf{A}_2)^2\} \lambda_{\max}}{\delta_{\min}^2} = o(1), \quad \text{and} \quad (C\text{-vi}') \quad \frac{\|\boldsymbol{\mu}_{12}\|^2 \lambda_{\max}}{\delta_{\min}^2} = o(1).$$

Proposition 3.1. *Assume that $\limsup_{p \rightarrow \infty} \lambda_{\max}(\mathbf{A}_i) < \infty$ for $i = 1, 2$. Then, (C-iv') and (C-vi') imply (C-iv) and (C-vi), respectively. Furthermore, if $\lambda(\mathbf{A}_i) \in (0, \infty)$ as $p \rightarrow \infty$ for $i = 1, 2$, and \mathbf{A}_i , $i = 1, 2$, are diagonal matrices such as in (I) to (III) in Section 1, (C-v') implies (C-v).*

Next, we consider the asymptotic normality of (1.2) under (A-ii). We assume the following condition instead of (C-vi) when $m \rightarrow \infty$ for $i = 1, 2$ ($i' \neq i$):

$$(C\text{-vii}) \quad \frac{\sum_{j=1}^{q_i} (\boldsymbol{\gamma}_{ij}^T \mathbf{A}_{i'} \boldsymbol{\mu}_{12})^4}{\delta_i^4} = o(1).$$

Note that $\sum_{j=1}^{q_i} (\boldsymbol{\gamma}_{is}^T \mathbf{A}_{i'} \boldsymbol{\mu}_{12})^4 \leq \sum_{j,j'=1}^{q_i} (\boldsymbol{\gamma}_{ij}^T \mathbf{A}_{i'} \boldsymbol{\mu}_{12})^2 (\boldsymbol{\gamma}_{ij'}^T \mathbf{A}_{i'} \boldsymbol{\mu}_{12})^2 = \Delta_{iA}^2$. Thus (C-vii) is milder than (C-vi).

Remark 3. *The condition in (C-vii) can be written as a condition concerning eigenvalues and eigenvectors. If $\boldsymbol{\Gamma}_i = \mathbf{H}_i \boldsymbol{\Lambda}_i^{1/2}$, $\mathbf{A}_i = \boldsymbol{\Sigma}_i^{-1}$, $i = 1, 2$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, it holds that $\sum_{j=1}^{q_i} \{\boldsymbol{\gamma}_{ij}^T \mathbf{A}_{i'} \boldsymbol{\mu}_{12}\}^4 = \sum_{j=1}^p \psi_j^2$ and $\Delta_{iA} = \boldsymbol{\mu}_{12}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_{12} = \sum_{j=1}^p \psi_j$, where $\psi_j = (\boldsymbol{\mu}_{12}^T \mathbf{h}_{ij})^2 / \lambda_{ij}$. Hence, the condition “ $\sum_{j=1}^p \psi_j^2 / (\sum_{j=1}^p \psi_j)^2 \rightarrow 0$ as $p \rightarrow \infty$ ” implies (C-vii).*

Now, we claim the asymptotic normality of (1.2) under (A-ii) as follows:

Theorem 3.2. *Assume (A-ii). Assume also (C-iv), (C-v) and (C-vii). Then, we have (3.1). Furthermore, for the classification rule by (1.3) with (1.2), we have (3.2).*

3.2 Bayes error rates

When considering Theorem 3.2 under the situation that

$$\text{tr}\{(\mathbf{\Sigma}_i \mathbf{A}_i)^2\}/n_i + \text{tr}(\mathbf{\Sigma}_i \mathbf{A}_{i'} \mathbf{\Sigma}_{i'} \mathbf{A}_{i'})/n_{i'} = o(\Delta_{iA}) \quad \text{as } m \rightarrow \infty \quad (3.3)$$

for $i = 1, 2$ ($i' \neq i$), one has (3.2) as

$$e(i) = \Phi\{-\Delta_i/(2\Delta_{iA}^{1/2})\} + o(1) \quad \text{as } m \rightarrow \infty \text{ for } i = 1, 2.$$

Note that $\delta_i/(2\Delta_{iA}^{1/2}) = 1 + o(1)$ under (3.3). If $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 (= \mathbf{\Sigma})$, the ratio $\Delta_i/\Delta_{iA}^{1/2}$ has a maximum when $\mathbf{A}_1 = \mathbf{A}_2 = \mathbf{\Sigma}^{-1}$. Then, the ratio becomes the Mahalanobis distance such as $\Delta_i/\Delta_{iA}^{1/2} = \Delta_{MD}^{1/2}$. The classification rule by (1.3) with (1.2) has an error rate converging to the Bayes error rate in the sense that $e(i) = \Phi(-\Delta_{MD}^{1/2}/2) + o(1)$ for $i = 1, 2$. On the other hand, if $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ and π_i s are Gaussian, under (C-iii) for (IV), the Bayes optimal classifier by (1.1) becomes as follows:

$$2(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_{i'}^{-1} \boldsymbol{\mu}_{12} + o_P(\Delta_{i(IV)}) > (-1)^i \Delta_{i(IV)}$$

when $\mathbf{x}_0 \in \pi_i$ ($i' \neq i$). Note that $\text{Var}\{(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_{i'}^{-1} \boldsymbol{\mu}_{12}\} = \boldsymbol{\mu}_{12}^T \mathbf{\Sigma}_{i'}^{-1} \mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1} \boldsymbol{\mu}_{12}$ (hereafter called $\Delta_{iA(IV)}$) when $\mathbf{x}_0 \in \pi_i$ ($i' \neq i$) and $\Delta_{iA(IV)}$ is the same as Δ_{iA} for (IV). Hence, $(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_{i'}^{-1} \boldsymbol{\mu}_{12}/\Delta_{iA(IV)}^{1/2}$ is distributed as $N(0, 1)$ when $\mathbf{x}_0 \in \pi_i : N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$. Then, the Bayes error rate becomes $e(i) = \Phi\{-\Delta_{i(IV)}/(2\Delta_{iA(IV)}^{1/2})\} + o(1)$ for $i = 1, 2$, under some conditions.

When considering Theorem 3.2 under the situation that

$$p/n_i + \text{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1})/n_{i'} = o(\Delta_{iA(IV)}) \quad \text{as } m \rightarrow \infty \quad (3.4)$$

for $i = 1, 2$ ($i' \neq i$), one can claim that the classification rule by (1.3) with (2.7) has the Bayes error rate asymptotically even when π_i s are non-Gaussian. Note that (3.4) is equivalent to (3.3) for (IV) and (3.4) usually holds when $n_{\min} \rightarrow \infty$ while p is fixed or $p \rightarrow \infty$ but $n_{\min}/p \rightarrow \infty$. If (3.4) is not met, the classifier by (2.7) is not optimal. We emphasize that (3.4) does not always hold for high-dimensional settings such as $n_{\min}/p \rightarrow 0$ or $n_{\min}/p \rightarrow c$ (> 0). For example, let us consider the setup of Figure 1. The condition “ $p/n_i = o(\Delta_{iA(IV)})$ ” is not met from the facts that $\Delta_{iA(IV)} = O(p^{2/3})$ and $n_1 = n_2 = o(p^{1/3})$, so that (3.4) does not hold. On the other hand, (C-iv) to (C-vi) hold, so that one can claim the asymptotic normality in Theorem 3.1. Note that (3.4) does not hold under (C-vi) for (IV). Thus the error rate of the classifier based on the Mahalanobis distance does not converge to the Bayes error rate when Theorem 3.1 is claimed. Such situations frequently occur in HDLSS settings such as $n_{\min}/p \rightarrow 0$. This is the reason why the classifier based on the Mahalanobis distance does not always give a preferable performance for high-dimensional data even when $n_{\min} \rightarrow \infty$, $\mathbf{\Sigma}_i$ s are known and π_i s are Gaussian.

3.3 Asymptotic normality for (I) to (IV)

We consider δ_i s for (I) to (IV). For (I), by putting $\mathbf{A}_i = \mathbf{I}_p$, $i = 1, 2$, one has δ_i ($i \neq i'$) as

$$\delta_i = 2 \left\{ \frac{\text{tr}(\mathbf{\Sigma}_i^2)}{n_i} + \frac{\text{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_{i'})}{n_{i'}} + \boldsymbol{\mu}_{12}^T \mathbf{\Sigma}_i \boldsymbol{\mu}_{12} \right\}^{1/2} \quad (\text{hereafter called } \delta_{i(I)}).$$

For (II), by putting $\mathbf{A}_i = \{p/\text{tr}(\mathbf{\Sigma}_i)\}\mathbf{I}_p, i = 1, 2$, it is given as

$$\delta_i = \frac{2p}{\text{tr}(\mathbf{\Sigma}_{i'})} \left\{ \frac{\delta_{i(I)}^2}{4} + \frac{\text{tr}(\mathbf{\Sigma}_i^2)}{n_i} \left(\frac{\text{tr}(\mathbf{\Sigma}_{i'})^2}{\text{tr}(\mathbf{\Sigma}_i)^2} - 1 \right) \right\}^{1/2} \text{ (hereafter called } \delta_{i(II)}).$$

For (III), by putting $\mathbf{A}_i = \mathbf{\Sigma}_{i(d)}^{-1}, i = 1, 2$, it is given as

$$\delta_i = 2 \left\{ \frac{\text{tr}\{(\mathbf{\Sigma}_i \mathbf{\Sigma}_{i(d)}^{-1})^2\}}{n_i} + \frac{\text{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_{i'(d)}^{-1} \mathbf{\Sigma}_{i'} \mathbf{\Sigma}_{i'(d)}^{-1})}{n_{i'}} + \boldsymbol{\mu}_{12}^T \mathbf{\Sigma}_{i'(d)}^{-1} \mathbf{\Sigma}_i \mathbf{\Sigma}_{i'(d)}^{-1} \boldsymbol{\mu}_{12} \right\}^{1/2} \\ \text{(hereafter called } \delta_{i(III)}).$$

For (IV), by putting $\mathbf{A}_i = \mathbf{\Sigma}_i^{-1}, i = 1, 2$, it is given as

$$\delta_i = 2 \left\{ \frac{p}{n_i} + \frac{\text{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1})}{n_{i'}} + \boldsymbol{\mu}_{12}^T \mathbf{\Sigma}_{i'}^{-1} \mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1} \boldsymbol{\mu}_{12} \right\}^{1/2} \text{ (hereafter called } \delta_{i(IV)}).$$

From Theorems 3.1, 3.2 and Proposition 3.1, we have the following result for (I) to (III).

Corollary 3.1. *Assume (C-iv'). Assume either (A-i) and (C-vi') or (A-ii) and (C-vii). Then, for the classification rule by (1.3) with (2.4), we have (3.2). Furthermore, under (C-v'), for the classification rule by (1.3) with (2.5) or (2.6), we have (3.2).*

Remark 4. *When $\text{tr}(\mathbf{\Sigma}_1)/\text{tr}(\mathbf{\Sigma}_2) \rightarrow 1$ as $p \rightarrow \infty$, it holds $\{\delta_{i(I)}p/\text{tr}(\mathbf{\Sigma}_{i'})\}/\delta_{i(II)} = 1 + o(1)$ ($i \neq i'$). Note that $\Delta_{i(II)}\text{tr}(\mathbf{\Sigma}_{i'})/p \geq \Delta_{(I)}$. It follows that $\Delta_{(I)}/\delta_{i(I)} \leq \Delta_{i(II)}/\delta_{i(II)}$ for sufficiently large p in (3.2).*

From Theorems 3.1 and 3.2 and Proposition 3.1, we have the following result for (IV).

Corollary 3.2. *Assume that (C-iv'), $\liminf_{p \rightarrow \infty} \lambda_{\min}(\mathbf{\Sigma}_i) > 0$ and $\text{tr}\{(\mathbf{I}_p - \mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1})^2\} = o(\delta_{\min(IV)}^2)$ for $i = 1, 2$ ($i' \neq i$), where $\delta_{\min(IV)} = \min\{\delta_{1(IV)}, \delta_{2(IV)}\}$. Assume either (A-i) and (C-vi') or (A-ii) and (C-vii). Then, for the classification rule by (1.3) with (2.7), we have (3.2).*

3.4 Comparisons of the classifiers

In this section, we investigate the performance of the classifier in (1.2) for (I) to (IV) by using the asymptotic normality. When $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, we consider (I), (III) and (IV) in the setup of Figure 1. Note that (I), (III) and (IV) satisfy (C-iv) to (C-vi) from the facts that $n_{\min} = o(p^{1/3})$, $\Delta_{iA} = O(\|\boldsymbol{\mu}_{12}\|^2) = O(p^{2/3})$, $\text{tr}(\mathbf{\Sigma}_i^2)/p \in (0, \infty)$ and $\text{tr}(\mathbf{\Sigma}_i^4) = o(p^2)$ as $p \rightarrow \infty$ for $i = 1, 2$. Thus, Theorem 3.1 holds for (I), (III) and (IV). We plotted the asymptotic error rates, $\Phi(-\Delta_{(I)}/\delta_{1(I)})$, $\Phi(-\Delta_{1(III)}/\delta_{1(III)})$ and $\Phi(-\Delta_{1(IV)}/\delta_{1(IV)})$ in Figure 3. From (3.2), we note that $e(1) - e(2) = o(1)$ when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Thus, the average error rate, $\bar{e} = \{e(1) + e(2)\}/2$, is regarded as an estimate of $e(1)$. We laid \bar{e} for (I), (III) and (IV) by borrowing from Figure 1. We observed that \bar{e} behaves very close to the asymptotic error rate as expected theoretically. We also plotted the Bayes error rate, $\Phi(-\Delta_{MD}^{1/2}/2)$. We observed that (IV) does not converge to the Bayes error rate when Theorem 3.1 is claimed. See Section 3.2 for the details. As for (I) and (III), the

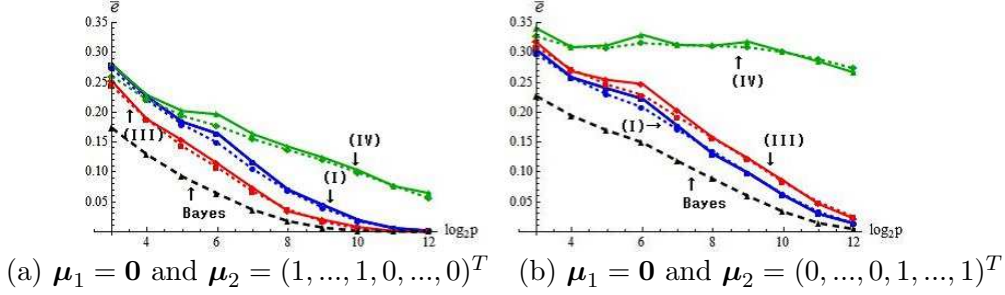


Figure 3: The asymptotic error rates (dashed lines) by $\Phi(-\Delta_{(I)}/\delta_{1(I)})$, $\Phi(-\Delta_{1(III)}/\delta_{1(III)})$ and $\Phi(-\Delta_{1(IV)}/\delta_{1(IV)})$, together with the corresponding \bar{e} (solid lines) by (2.4), (2.6) and (2.7) in the setup of Figure 1. The Bayes optimal error rate was given by $\Phi(-\Delta_{MD}^{1/2}/2)$.

difference of the performances depends on the configuration of μ_{ij} s and $\sigma_{i(j)}$ s. When p is sufficiently large, we note that $\Delta_{(I)} = \sum_{j=1}^p \mu_{12j}^2 < \Delta_{1(III)} = \sum_{j=1}^p \mu_{12j}^2 / \sigma_{2(j)}$ for (a) and $\Delta_{(I)} > \Delta_{1(III)}$ for (b) because $\sigma_{2(j)} = 0.5 + j/(p+1)$, $j = 1, \dots, p$ both for (a) and (b). It follows that $\Delta_{(I)}/\delta_{i(I)} < \Delta_{i(III)}/\delta_{i(III)}$ for (a) and $\Delta_{(I)}/\delta_{i(I)} > \Delta_{i(III)}/\delta_{i(III)}$ for (b). Thus (III) is better than (I) for (a), on the other hand, they trade places for (b).

When $\Sigma_1 \neq \Sigma_2$, (II), (III) and (IV) draw information about heteroscedasticity through the difference of $\text{tr}(\Sigma_i)$ s, $\Sigma_{i(d)}$ s or Σ_i s, respectively. We consider them in the setup of Figure 2. For (c), note that $\Delta_{(I)} = 0$ but $\Delta_{i(II)} = \Delta_{i(III)} = \Delta_{i(IV)} > cp$ for some constant $c > 0$. (II), (III) and (IV) hold the consistency property even when n_i s are fixed because (C-i) to (C-iii) are satisfied. Actually, in Figure 2, we observed that the three classifiers gave preferable performances by using the difference of $\text{tr}(\Sigma_i)$ s, $\Sigma_{i(d)}$ s or Σ_i s as p increases. For (d), note that the difference of $\text{tr}(\Sigma_i)$ s is smaller than that for (c). Actually, in Figure 2, we observed that (II) gives a worse performance for (d) compared to (c). On the other hand, (III) gave a better performance compared to (II) because $\Delta_{i(III)}$ is sufficiently larger than $\Delta_{i(II)}$ for (d) when p is large. (IV) draws information about heteroscedasticity from the difference of the covariance matrices themselves, so that it gave the best performance in this case. However, we note that it is quite difficult to estimate Σ_i^{-1} s feasibly for high-dimensional data. See Section 5.2 for the details.

4 Estimation of the quadratic classifier

We denote an estimator of \mathbf{A}_i by $\hat{\mathbf{A}}_i$. We consider estimating the quadratic classifier by $W_i(\hat{\mathbf{A}}_i)$.

4.1 Preliminary

Let $\|\mathbf{M}\| = \lambda_{\max}^{1/2}(\mathbf{M}^T \mathbf{M})$ for any square matrix \mathbf{M} . Let κ be a constant such as $\kappa = \Delta_{\min}$ or $\kappa = \delta_{\min}$. We consider the following condition for $\hat{\mathbf{A}}_i$ s under (\star):

(C-viii) $p\|\hat{\mathbf{A}}_i - \mathbf{A}_i\| = o_P(\kappa)$ for $i = 1, 2$.

Proposition 4.1. *Assume (C-viii). Assume also that $\lambda(\mathbf{A}_i) \in (0, \infty)$ as $p \rightarrow \infty$ for $i = 1, 2$. Then, we have that*

$$W_1(\hat{\mathbf{A}}_1) - W_2(\hat{\mathbf{A}}_2) = W_1(\mathbf{A}_1) - W_2(\mathbf{A}_2) + o_P(\kappa) \quad (4.1)$$

under (\star) when $\mathbf{x}_0 \in \pi_i$ for $i = 1, 2$.

When one chooses \mathbf{A}_i s as $\mathbf{A}_1 = \mathbf{A}_2 (= \mathbf{A})$, $W(\hat{\mathbf{A}})$ gives a linear classifier. We consider the following condition for $\hat{\mathbf{A}}$ under (\star) :

$$(\mathbf{C}\text{-ix}) \quad (p/n_{\min}^{1/2} + p^{1/2}\|\boldsymbol{\mu}_{12}\|)\|\hat{\mathbf{A}} - \mathbf{A}\| = o_P(\kappa).$$

We have the following result.

Proposition 4.2. *Assume (C-ix). Then, we have (4.1).*

We note that (C-ix) is milder than (C-viii) from the fact that $\|\boldsymbol{\mu}_{12}\| = O(p^{1/2})$. Hence, we recommend to use a linear classifier such as (2.4) or (4.5). The quadratic classifiers should be used when the difference of covariance matrices is considerably large. See Section 4.3 for the details.

4.2 Quadratic classifier by $\hat{\mathbf{A}}_i = \{p/\text{tr}(\mathbf{S}_i)\}\mathbf{I}_p$

We consider the classifier by

$$W_i(\{p/\text{tr}(\mathbf{S}_{in_i})\}\mathbf{I}_p) = \frac{p\|\mathbf{x}_0 - \bar{\mathbf{x}}_{in_i}\|^2}{\text{tr}(\mathbf{S}_{in_i})} - \frac{p}{n_i} + p \log\{\text{tr}(\mathbf{S}_{in_i})/p\}. \quad (4.2)$$

Note that $\delta_i = \delta_{i(II)}$, $\Delta_i = \Delta_{i(II)}$ and $\mathbf{A}_i = \{p/\text{tr}(\boldsymbol{\Sigma}_i)\}\mathbf{I}_p$. Here, $\lambda(\mathbf{A}_i) \in (0, \infty)$ as $p \rightarrow \infty$ for $i = 1, 2$, and (C-viii) naturally holds. From Corollary 2.1 and Proposition 4.1, we have the following result.

Corollary 4.1. *Assume (A-i). Assume also (C-i') and (C-ii'). Then, for the classification rule by (1.3) with (4.2), we have (2.2).*

The classifier by (4.2) is equivalent to the geometric classifier by Aoshima and Yata (2011). Hereafter, we call the classifier by (4.2) the “geometrical quadratic discriminant analysis (GQDA)”. Similar to Section 2.2, we have (2.2) for GQDA under (A-i) and (2.3) even when n_{\min} is fixed. If one can assume that $\liminf_{p \rightarrow \infty} |\text{tr}(\boldsymbol{\Sigma}_1)/\text{tr}(\boldsymbol{\Sigma}_2) - 1| > 0$, we have (2.2) for GQDA under (A-i) and (2.8) even when n_{\min} is fixed and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. As for the asymptotic normality, by combining Corollary 3.1 with Lemma B.3 given in Appendix B, we have the following result.

Corollary 4.2. *Assume (C-iv') and (C-v'). Assume either (A-i) and (C-vi') or (A-ii) and (C-vii). Then, for the classification rule by (1.3) with (4.2), we have (3.2) under $(\text{tr}(\boldsymbol{\Sigma}_1)/\text{tr}(\boldsymbol{\Sigma}_2) - 1)^2 \text{tr}(\boldsymbol{\Sigma}_{\max}^2) = o(n_{\min} \delta_{\min(II)}^2)$ as $m \rightarrow \infty$, where $\delta_{\min(II)} = \min\{\delta_{1(II)}, \delta_{2(II)}\}$.*

Now, we compare DBDA with GQDA. We have that

$$\begin{aligned}\hat{\Delta}_{(I)} &= \|\bar{\mathbf{x}}_{1n_1} - \bar{\mathbf{x}}_{2n_2}\|^2 - \text{tr}(\mathbf{S}_{1n_1})/n_1 - \text{tr}(\mathbf{S}_{2n_2})/n_2 \quad \text{and} \\ \hat{\Delta}_{i(II)} &= \frac{p}{\text{tr}(\mathbf{S}_{i'n_{i'}})} \left[\hat{\Delta}_{(I)} + \text{tr}(\mathbf{S}_{in_i}) - \text{tr}(\mathbf{S}_{i'n_{i'}}) + \text{tr}(\mathbf{S}_{i'n_{i'}}) \log \left\{ \frac{\text{tr}(\mathbf{S}_{i'n_{i'}})}{\text{tr}(\mathbf{S}_{in_i})} \right\} \right]\end{aligned}$$

for $i = 1, 2$ ($i' \neq i$). We note that $E(\hat{\Delta}_{(I)}) = \Delta_{(I)}$. From (3.2) and Remark 4, if $\hat{\Delta}_{i(II)} \text{tr}(\mathbf{S}_{i'n_{i'}})/p$ is sufficiently larger than $\hat{\Delta}_{(I)}$ for some i , we recommend to use GQDA. Otherwise one may use DBDA free from (A-i). See Corollary 2.1 for the details.

4.3 Quadratic classifier by $\hat{\mathbf{A}}_i = \mathbf{S}_{in_i(d)}^{-1}$

Let $\mathbf{S}_{in_i(d)} = \text{diag}(s_{in_i(1)}, \dots, s_{in_i(p)})$ for $i = 1, 2$. We consider the classifier by

$$W_i(\mathbf{S}_{in_i(d)}^{-1}) = \sum_{j=1}^p \left(\frac{(x_{0j} - \bar{x}_{ijn_i})^2}{s_{in_i(j)}} - \frac{1}{n_i} + \log s_{in_i(j)} \right). \quad (4.3)$$

Note that $\delta_i = \delta_{i(III)}$, $\Delta_i = \Delta_{i(III)}$ and $\mathbf{A}_i = \mathbf{\Sigma}_{i(d)}^{-1}$. Dudoit, Fridlyand and Speed (2002) considered the quadratic classifier without the bias correction term. That was called the diagonal quadratic discriminant analysis (DQDA). Hereafter, we call the classifier by (4.3) ‘‘DQDA-bc’’. Let $\eta_{i(j)} = \text{Var}\{(x_{ijk} - \mu_{ij})^2\}$ for $i = 1, 2$, and $j = 1, \dots, p$ ($k = 1, \dots, n_i$). Since $\hat{\mathbf{A}}_i = \mathbf{S}_{in_i(d)}^{-1}$ does not satisfy (C-viii) in that shape, we consider the following assumption:

(A-iii) $\eta_{i(j)} \in (0, \infty)$ as $p \rightarrow \infty$ and $\limsup_{p \rightarrow \infty} E\{\exp(t_{ij}|x_{ijk} - \mu_{ij}|^2/\eta_{i(j)}^{1/2})\} < \infty$ for some $t_{ij} > 0$, $i = 1, 2$, and $j = 1, \dots, p$ ($k = 1, \dots, n_i$).

Note that (A-iii) holds when π_i has $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for $i = 1, 2$. From Corollary 2.1 and Proposition 4.1, we have the following result.

Corollary 4.3. *Assume (A-i) and (A-iii). Assume also (C-ii’). Then, for the classification rule by (1.3) with (4.3), we have (2.2) under the condition that*

$$\frac{p^2 \log p}{n_{\min} \Delta_{\min(III)}^2} = o(1). \quad (4.4)$$

Note that (C-i’) holds under (4.4). From the fact that $\Delta_{i(III)} = O(p)$, it follows that $n_{\min}^{-1} \log p = o(1)$ under (4.4). Similar to Section 2.2, if one can assume that $\liminf_{p \rightarrow \infty} \|\boldsymbol{\mu}_{12}\|^2/p > 0$ or $\liminf_{p \rightarrow \infty} \sum_{j=1}^p |\sigma_{1(j)}/\sigma_{2(j)} - 1|/p > 0$, DQDA-bc holds (2.2) under (A-i), (A-iii), (2.8) and $n_{\min}^{-1} \log p = o(1)$. When $\Delta_{\min(III)}$ is not sufficiently large, say $\Delta_{\min(III)} = O(p^{1/2})$, we can claim Corollary 4.3 in high-dimension, large-sample-size settings such as $n_{\min}/p \rightarrow \infty$. In Section 5, we shall provide a DQDA type classifier by feature selection and show that it has the consistency property even when $n_{\min}/p \rightarrow 0$ and $\Delta_{\min(III)}$ is not sufficiently large.

Next, we consider the pooled sample diagonal matrix,

$$\mathbf{S}_{n(d)} = \frac{\sum_{i=1}^2 (n_i - 1) \mathbf{S}_{in_i(d)}}{\sum_{i=1}^2 n_i - 2}.$$

Note that $E(\mathbf{S}_{n(d)}) = \sum_{i=1}^2 (n_i - 1) \boldsymbol{\Sigma}_{i(d)} / (\sum_{i=1}^2 n_i - 2)$ (hereafter called $\boldsymbol{\Sigma}_{(d)}$). When $\boldsymbol{\Sigma}_{1(d)} = \boldsymbol{\Sigma}_{2(d)}$, it follows that $\boldsymbol{\Sigma}_{(d)} = \boldsymbol{\Sigma}_{i(d)}$, $i = 1, 2$. Let us write $\mathbf{S}_{n(d)} = \text{diag}(s_{n(1)}, \dots, s_{n(p)})$ and $\boldsymbol{\Sigma}_{(d)} = \text{diag}(\sigma_{(1)}, \dots, \sigma_{(p)})$. We consider the classifier by

$$W_i(\mathbf{S}_{n(d)}^{-1}) = \sum_{j=1}^p \left(\frac{(x_{0j} - \bar{x}_{ijn_i})^2}{s_{n(j)}} - \frac{s_{in_i(j)}}{n_i s_{n(j)}} \right). \quad (4.5)$$

We note that the classification rule by (1.3) with (4.5) becomes a linear classifier. Bickel and Levina (2004) and Dudoit, Fridlyand and Speed (2002) considered the linear classifier without the bias correction term. That was called the diagonal linear discriminant analysis (DLDA). Hereafter, we call the classifier by (4.5) ‘‘DLDA-bc’’. Although Huang, Tong and Zhao (2010) gave bias corrected versions of DLDA and DQDA, they considered a bias correction only when π_i s are Gaussian. We note that $\Delta_1 = \Delta_2 = \sum_{j=1}^p \mu_{12j}^2 / \sigma_{(j)}$ (hereafter called $\Delta_{(III')}$) and $\mathbf{A}_1 = \mathbf{A}_2 = \boldsymbol{\Sigma}_{(d)}^{-1}$. Then, by combining Theorem 2.1 with Propositions 2.1 and 4.2, we have the following result.

Corollary 4.4. *Assume (A-iii). Assume also (C-i') and (C-ii'). Then, for the classification rule by (1.3) with (4.5), we have (2.2) under the condition that*

$$\frac{p \log p}{n_{\min} \Delta_{(III')}} = o(1). \quad (4.6)$$

Under $n_{\min}^{-1} \log p = o(1)$, one may claim that (4.6) is milder than (4.4) if $\Delta_{\min(III)}$ and $\Delta_{(III')}$ are of the same order. Hence, we recommend to use DQDA-bc when $\Delta_{\min(III)}$ is considerably larger than $\Delta_{(III')}$. Otherwise one may use DLDA-bc even when $\boldsymbol{\Sigma}_{i(d)}$ s are not common. We shall improve DQDA-bc by feature selection in Section 5.

4.4 Quadratic classifier by $\hat{\mathbf{A}}_i = \mathbf{S}_{in_i}^{-1}$

In this section, we consider high-dimension, large-sample-size situations such as $n_{\min}/p \rightarrow \infty$ as $p \rightarrow \infty$ and discuss the classifier by

$$W_i(\mathbf{S}_{in_i}^{-1}) = (\mathbf{x}_0 - \bar{\mathbf{x}}_{in_i})^T \mathbf{S}_{in_i}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_{in_i}) - p/n_i + \log |\mathbf{S}_{in_i}|. \quad (4.7)$$

Note that $\delta_i = \delta_{i(IV)}$, $\Delta_i = \Delta_{i(IV)}$ and $\mathbf{A}_i = \boldsymbol{\Sigma}_i^{-1}$. Let $\eta_{i(rs)} = \text{Var}\{(x_{irk} - \mu_{ir})(x_{isk} - \mu_{is})\}$ for $i = 1, 2$, and $r, s = 1, \dots, p$ ($k = 1, \dots, n_i$). From Theorem 2.1 and Proposition 4.1, we have the following result.

Corollary 4.5. *Assume (A-i) and (A-iii). Assume also $\lambda(\boldsymbol{\Sigma}_i) \in (0, \infty)$ as $p \rightarrow \infty$ and $\liminf_{p \rightarrow \infty} \eta_{i(rs)} > 0$ for all r, s ; $i = 1, 2$. Then, for the classification rule by (1.3) with (4.7), we have (2.2) under the conditions that $p^{1/2}/\Delta_{\min(IV)} = o(1)$ and*

$$\frac{p^4 \log p}{n_{\min} \Delta_{\min(IV)}^2} = o(1). \quad (4.8)$$

From the fact that $\Delta_{i(IV)} = O(p)$ when $\lambda(\Sigma_i) \in (0, \infty)$ as $p \rightarrow \infty$ for $i = 1, 2$, it follows that $n_{\min}^{-1} p^2 \log p = o(1)$ under (4.8). Thus, the classification rule by (1.3) with (4.7) can claim the consistency property when $n_{\min}^{-1} p^2 \log p = o(1)$. However, the condition “ $n_{\min}^{-1} p^2 \log p = o(1)$ ” is quite strict for high-dimensional data. In Section 5, we shall discuss a classifier by sparse inverse covariance matrix estimation when $n_{\min}/p \rightarrow 0$.

5 Quadratic classifiers by feature selection and sparse inverse covariance matrix estimation

In this section, we propose a new quadratic classifier by feature selection for (4.3) and discuss a quadratic classifier by sparse inverse covariance matrix estimation for (4.7).

5.1 Quadratic classifier after feature selection

We consider applying a variable selection procedure to classification. Fan and Fan (2008) proposed the feature annealed independent rule based on the difference of mean vectors. However, we give a different type of feature selection by using both the differences of mean vectors and covariance matrices. We have that

$$\Delta_{1(III)} + \Delta_{2(III)} = \sum_{j=1}^p \left(\frac{\mu_{12j}^2 + \sigma_{1(j)}}{\sigma_{2(j)}} + \frac{\mu_{12j}^2 + \sigma_{2(j)}}{\sigma_{1(j)}} - 2 \right).$$

Let $\theta_j = (\mu_{12j}^2 + \sigma_{1(j)})/(2\sigma_{2(j)}) + (\mu_{12j}^2 + \sigma_{2(j)})/(2\sigma_{1(j)}) - 1$ for $j = 1, \dots, p$. Note that $\Delta_{1(III)} + \Delta_{2(III)} = 2 \sum_{j=1}^p \theta_j$. Also, note that $\theta_j > 0$ when $\mu_{1j} \neq \mu_{2j}$ or $\sigma_{1(j)} \neq \sigma_{2(j)}$. Now, we give an estimator of θ_j ($j = 1, \dots, p$) by

$$\hat{\theta}_j = \frac{(\bar{x}_{1jn_1} - \bar{x}_{2jn_2})^2 + s_{1n_1(j)}}{2s_{2n_2(j)}} + \frac{(\bar{x}_{1jn_1} - \bar{x}_{2jn_2})^2 + s_{2n_2(j)}}{2s_{1n_1(j)}} - 1.$$

Then, we have the following result.

Theorem 5.1. *Assume (A-iii). Assume also $n_{\min}^{-1} \log p = o(1)$. Then, we have that as $p \rightarrow \infty$*

$$\max_{j=1, \dots, p} |\hat{\theta}_j - \theta_j| = O_P\{(n_{\min}^{-1} \log p)^{1/2}\}.$$

Let $\mathbf{D} = \{j \mid \theta_j > 0 \text{ for } j = 1, \dots, p\}$ and $p_* = \#\mathbf{D}$, where $\#\mathbf{S}$ denotes the number of elements in a set \mathbf{S} . Let $\xi = (n_{\min}^{-1} \log p)^{1/2}$. We select a set of significant variables by

$$\hat{\mathbf{D}} = \{j \mid \hat{\theta}_j > \xi^\gamma \text{ for } j = 1, \dots, p\}, \quad (5.1)$$

where $\gamma \in (0, 1)$ is a chosen constant. Then, from Theorem 5.1, we have the following result.

Corollary 5.1. *Assume (A-iii) and $n_{\min}^{-1} \log p = o(1)$. Assume also $\liminf_{p \rightarrow \infty} \theta_j > 0$ for all $j \in \mathbf{D}$. Then, we have that $P(\mathbf{D} = \hat{\mathbf{D}}) \rightarrow 1$ as $p \rightarrow \infty$.*

Remark 5. *As for l (≥ 3)-class classification, one may consider $\hat{\theta}_j$ such as $\hat{\theta}_j = \sum_{i \neq i'}^k \{(\bar{x}_{ijn_i} - \bar{x}_{i'jn_{i'}})^2 + s_{in_i(j)}\} / \{k(k-1)s_{i'n_{i'}(j)}\} - 1$ for $j = 1, \dots, p$.*

Now, we consider a classifier using only the variables in $\hat{\mathbf{D}}$. We define the classifier by

$$W_i(\mathbf{S}_{in_i(d)}^{-1})_{FS} = \sum_{j \in \hat{\mathbf{D}}} \left(\frac{(x_{0j} - \bar{x}_{ijn_i})^2}{s_{in_i(j)}} - \frac{1}{n_i} + \log s_{in_i(j)} \right) \quad (5.2)$$

for $i = 1, 2$. We consider the classification rule by (1.3) with (5.2). We call this feature selected DQDA “FS-DQDA”. Let us write that $\mathbf{x}_{i*k} = (x_{ij_1k}, \dots, x_{ij_{p_*}k})^T$ for all i, k , where $\mathbf{D} = \{j_1, \dots, j_{p_*}\}$. Let $\boldsymbol{\Sigma}_{i*} = \text{Var}(\mathbf{x}_{i*k})$ for $i = 1, 2$ ($k = 1, \dots, n_i$). Then, from Theorem 2.1 and Corollary 5.1, we have the following result.

Corollary 5.2. *Assume (A-i) and (A-iii). Assume also $\lambda_{\max}(\boldsymbol{\Sigma}_{i*}) = o(p_*)$ for $i = 1, 2$, and $\liminf_{p \rightarrow \infty} \theta_j > 0$ for all $j \in \mathbf{D}$. Then, for the classification rule by (1.3) with (5.2), we have (2.2) under $n_{\min}^{-1} \log p = o(1)$.*

By comparing Corollary 5.2 with 4.3, note that the condition “ $n_{\min}^{-1} \log p = o(1)$ ” is much milder than (4.4). Thus we recommend FS-DQDA more than DQDA-bc (or the original DQDA). For a choice of $\gamma \in (0, 1)$ in (5.1), we recommend applying cross-validation procedures or choosing a constant such as $\gamma = 0.5$ because Corollary 5.2 is claimed for any $\gamma \in (0, 1)$. In addition, we emphasize that the computational cost of FS-DQDA is quite low even when $p \geq 10,000$.

5.2 Quadratic classifier by sparse inverse covariance matrix estimation

We consider applying a sparse estimation of inverse covariance matrices to classification. Bickel and Levina (2008b) gave a sparse estimator of $\boldsymbol{\Sigma}_i^{-1}$. Let $\sigma_{i(st)}$ be the (s, t) element of $\boldsymbol{\Sigma}_i$ for $s, t = 1, \dots, p$ ($i = 1, 2$). A sparsity measure of $\boldsymbol{\Sigma}_i$ ($i = 1, 2$) is given by $c_{p, h_i} = \max_{1 \leq t \leq p} \sum_{s=1}^p |\sigma_{i(st)}|^{h_i}$ for $0 \leq h_i < 1$, where 0^0 is defined to be 0. Note that $\lambda_{\max}(\boldsymbol{\Sigma}_i) \leq M c_{p, h_i}$ for some constant $M > 0$. If c_{p, h_i} is much smaller than p for a constant $h_i \in [0, 1)$, $\boldsymbol{\Sigma}_i$ is considered as sparse in the sense that many elements of $\boldsymbol{\Sigma}_i$ are very small. See Section 3 in Shao et al. (2011) for the details. Let $I(\cdot)$ be the indicator function. A thresholding operator is defined by $T_\tau(\mathbf{M}) = [m_{st} I(|m_{st}| \geq \tau)]$ for any $\tau > 0$ and any symmetric matrix $\mathbf{M} = [m_{st}]$. Let $\tau_{n_i} = M'(n_i^{-1} \log p)^{1/2}$ for some constant $M' > 0$. Then, Bickel and Levina (2008b) gave the following result.

Theorem 5.2. *Assume (A-iii), $n_i^{-1} \log p = o(1)$ and $\liminf_{p \rightarrow \infty} \lambda_{\min}(\boldsymbol{\Sigma}_i) > 0$. For a sufficiently large $M'(> 0)$, it holds that as $p \rightarrow \infty$*

$$\| \{T_{\tau_{n_i}}(\mathbf{S}_{in_i})\}^{-1} - \boldsymbol{\Sigma}_i^{-1} \| = O_P \left(c_{p, h_i} (n_i^{-1} \log p)^{(1-h_i)/2} \right).$$

Remark 6. *Theorem 5.2 is obtained by Theorem 1 and Section 2.3 in Bickel and Levina (2008b).*

We use $\hat{\mathbf{A}}_i = \{T_{\tau_{n_i}}(\mathbf{S}_{in_i})\}^{-1}$ as an estimator of $\boldsymbol{\Sigma}_i^{-1}$ and consider the classifier by $W_i(\{T_{\tau_{n_i}}(\mathbf{S}_{in_i})\}^{-1})$. By combining Theorem 5.2 and Proposition 4.1, if it holds that $\lambda(\boldsymbol{\Sigma}_i) \in (0, \infty)$ as $p \rightarrow \infty$ and

$$\frac{p c_{p, h_i} (n_i^{-1} \log p)^{(1-h_i)/2}}{\Delta_{\min}(\mathbf{IV})} = o_P(1), \quad (5.3)$$

the classification rule by (1.3) with $W_i(\{T_{\tau_{n_i}}(\mathbf{S}_{in_i})\}^{-1})$ has (2.2) under some regularity conditions. When $\mathbf{\Sigma}_i$ s are sparse as $c_{p,h_i} = O(1)$ for some $h_i (i = 1, 2)$ and $\liminf_{p \rightarrow \infty} \Delta_{\min}(IV)/p > 0$, (5.3) holds in HDLSS situations such as $n_{\min}^{-1} \log p = o(1)$. Shao et al. (2011) and Li and Shao (2015) considered a linear and a quadratic classifier by the sparse estimation of $\mathbf{\Sigma}_i^{-1}$ s under some sparsity conditions. On the other hand, Cai, Liu and Luo (2011) gave the constrained ℓ_1 -minimization for inverse matrix estimation (CLIME). One may apply the CLIME to the classification rule by (1.3). However, one should note that the computational cost for the sparse estimation of $\mathbf{\Sigma}_i^{-1}$ s is extremely high even when $p \approx 1,000$. It is quite unrealistic to apply the estimation to classification when p is very high as $p \geq 10,000$. Also, the sparsity condition “ $\lambda(\mathbf{\Sigma}_i) \in (0, \infty)$ as $p \rightarrow \infty$ ” is quite severe for high-dimensional data. In actual data analyses, we often encounter the situation that $\lambda_{ij} \rightarrow \infty$ as $p \rightarrow \infty$ for the first several j s. See Yata and Aoshima (2013) for the details.

5.3 Simulation

We used computer simulations to compare the performance of the classifiers: DBDA by (2.4), GQDA by (4.2), DLDA-bc by (4.5), DQDA-bc by (4.3) and FS-DQDA by (5.2). We did not compare the classifiers with the one given by sparse estimation of $\mathbf{\Sigma}_i^{-1}$ s such as $W_i(\{T_{\tau_{n_i}}(\mathbf{S}_{in_i})\}^{-1})$ in Section 5.2 because the computational cost of the sparse estimation is very high when p is large. Thus we considered the classifier by (2.7) instead of using the sparse estimation, provided that $\mathbf{\Sigma}_i$ s were known. We set $\gamma = 0.5$ in (5.1). We considered $p_* = \lceil p^{1/2} \rceil$. We generated $\mathbf{x}_{ik} - \boldsymbol{\mu}_i$, $k = 1, 2, \dots$, ($i = 1, 2$) independently from (i) $N_p(\mathbf{0}, \mathbf{\Sigma}_i)$ or (ii) a p -variate t -distribution, $t_p(\mathbf{0}, \mathbf{\Sigma}_i, \nu)$ with mean zero, covariance matrix $\mathbf{\Sigma}_i$ and degrees of freedom ν . We set $p = 2^s$, $s = 3, \dots, 10$ for (i), and $p = 500$ and $\nu = 4s$, $s = 1, \dots, 8$ for (ii). We set $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = (0, \dots, 0, 1, \dots, 1)^T$ whose last p_* elements are 1 and $\mathbf{\Sigma}_1 = \mathbf{B}_1(0.3^{|i-j|^{1/3}})\mathbf{B}_1$, where \mathbf{B}_1 is defined in Section 1. Let $\mathbf{B}_2 = \text{diag}(1, \dots, 1, 2^{1/2}, \dots, 2^{1/2})$ whose last p_* diagonal elements are $2^{1/2}$. We considered four cases:

- (a) $n_1 = 10$, $n_2 = 20$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_1$ for (i) $N_p(\mathbf{0}, \mathbf{\Sigma}_i)$;
- (b) $n_1 = \lceil (\log p)^2 \rceil$, $n_2 = 2n_1$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_1$ for (i) $N_p(\mathbf{0}, \mathbf{\Sigma}_i)$;
- (c) $n_1 = \lceil (\log p)^2 \rceil$, $n_2 = 2n_1$ and $\mathbf{\Sigma}_2 = \mathbf{B}_2\mathbf{\Sigma}_1\mathbf{B}_2$ for (i) $N_p(\mathbf{0}, \mathbf{\Sigma}_i)$;
- and (d) $n_1 = \lceil (\log p)^2 \rceil$, $n_2 = 2n_1$ and $\mathbf{\Sigma}_2 = \mathbf{B}_2\mathbf{\Sigma}_1\mathbf{B}_2$ for (ii) $t_p(\mathbf{0}, \mathbf{\Sigma}_i, \nu)$.

It holds that $n_{\min}^{-1} \log p = o(1)$ for (b), (c) and (d), $\liminf_{p \rightarrow \infty} \Delta_{\min}/p_* > 0$ for (a) to (d), and $\liminf_{p \rightarrow \infty} |\text{tr}(\mathbf{\Sigma}_1) - \text{tr}(\mathbf{\Sigma}_2)|/p_* > 0$ for (c) and (d). Similar to Section 1, we calculated the average error rate, \bar{e} , by 2000 replications and plotted the results in Figure 4 (a) to (d).

We observed from (a) in Figure 4 that DBDA and GQDA give preferable performances when n_i s are fixed. DLDA-bc gave a moderate performance because $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. However, the other classifiers did not give preferable performances when p is large. This is probably due to the consistency property of those classifiers (except (2.7)) which is claimed under at least $n_{\min}^{-1} \log p = o(1)$. Actually, as for (b), the other classifiers gave moderate performances because $n_{\min}^{-1} \log p = o(1)$. Thus we do not recommend to use quadratic classifiers including all the elements (or the diagonal elements) of sample covariance matrices, such as DQDA-bc and FS-DQDA, when the condition “ $n_{\min}^{-1} \log p = o(1)$ ” is not satisfied. When $n_{\min}^{-1} \log p \neq o(1)$ or n_i s are fixed, we recommend to use DBDA and

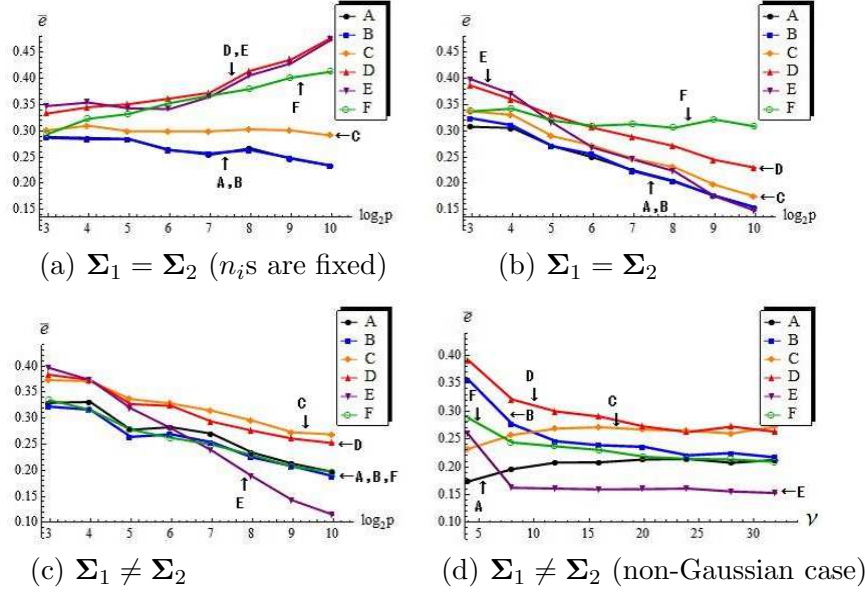


Figure 4: The average error rates of the classifiers: A: DBDA, B: GQDA, C: DLDA-bc, D: DQDA-bc, E: FS-DQDA, and F: the classifier by (2.7).

GQDA. On the other hand, FS-DQDA gave a good performance for (c) as p increases because the difference of the covariance matrices becomes large as p increases. We note that from Corollary 5.2 FS-DQDA holds the consistency property for (c). However, DQDA-bc did not give a preferable performance because $\Delta_{\min}(III) = O(p^{1/2})$, so that DQDA-bc does not hold the consistency property from Corollary 4.3. We note that $\Sigma_1 \neq \Sigma_2$ but $\Delta_{(I)}/\delta_{i(I)} \approx \Delta_{i(II)}/\delta_{i(II)}$ for (c). Thus GQDA gave a similar performance to DBDA for (c). As for (d), DBDA gave a preferable performance even when ν is small because DBDA holds the consistency property without (A-i). The other classifiers did not give preferable performances when ν is small. However, these classifiers gave moderate performances when ν becomes large because $t_p(\mathbf{0}, \Sigma_i, \nu) \Rightarrow N_p(\mathbf{0}, \Sigma_i)$ as $\nu \rightarrow \infty$. Especially, FS-DQDA gave a good performance when ν is not small. This is probably because FS-DQDA has smaller variance by feature selection, such as $p_*/p \rightarrow 0$, compared to the other classifiers.

Throughout the simulations, the classifier by (2.7) did not give preferable performances in spite that Σ_i s are known. See Section 3.2 for theoretical reasons. Therefore, it is likely that the classifier by $W_i(\{T_{\tau_{n_i}}(\mathbf{S}_{in_i})\}^{-1})$ gives poor performances for the high-dimensional settings.

6 Example: Leukemia data sets

We first analyzed gene expression data given by Golub et al. (1999) in which the data set consists of 7129 ($= p$) genes and 72 samples. We had 2 classes of leukemia subtypes, that is, π_1 : acute lymphoblastic leukemia (ALL) (47 samples) and π_2 : acute myeloid leukemia (AML) (25 samples). The data set consisted of two sets as 38 training samples (ALL: 27 samples and AML: 11 samples) and 34 test samples (ALL: 20 samples and AML: 14 samples). Note that $\mathbf{S}_{1n_1(d)} = \mathbf{S}_{2n_2(d)}$ if each sample has unit variance. Thus we did not

standardize each sample so as to have unit variance.

First, we checked several sparsity conditions. We standardized each sample by $\mathbf{x}_{ik} / \{\sum_{l=1}^2 \text{tr}(\mathbf{S}_{l n_l}) / (2p)\}^{1/2}$ for all i, k , so that $\text{tr}(\mathbf{S}_{1 n_1})/2 + \text{tr}(\mathbf{S}_{2 n_2})/2 = p$. By using all the samples (i.e., 72 samples), we calculated that

$$\hat{\Delta}_{(I)} = 2060 \quad (= 0.289p), \quad (6.1)$$

where $\hat{\Delta}_{(I)}$ is given in Section 4.2. Note that $E(\hat{\Delta}_{(I)}) = \|\boldsymbol{\mu}_{12}\|^2$. From this observation, we concluded that $\boldsymbol{\mu}_{12}$ is non-sparse. Next, we considered an estimator of $\|\boldsymbol{\Sigma}_{12}\|_F^2 = \sum_{i=1}^2 \text{tr}(\boldsymbol{\Sigma}_i^2) - 2\text{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)$ by $\hat{\Delta}_{\Sigma} = \sum_{i=1}^2 W_{in_i} - 2\text{tr}(\mathbf{S}_{1 n_1} \mathbf{S}_{2 n_2})$ having W_{in_i} s defined by (16) in Aoshima and Yata (2014). Here, W_{in_i} is an unbiased estimator of $\text{tr}(\boldsymbol{\Sigma}_i^2)$, so that $E(\hat{\Delta}_{\Sigma}) = \|\boldsymbol{\Sigma}_{12}\|_F^2$. We calculated that

$$\hat{\Delta}_{\Sigma} = 9.77 \times 10^5 \quad (= 137p).$$

From this observation, we concluded that $\boldsymbol{\Sigma}_{12}$ is non-sparse. Therefore, the Bayes error rates of this data set are probably close to 0. Also, we calculated

$$(\hat{\lambda}_{\max}(\boldsymbol{\Sigma}_1), \hat{\lambda}_{\max}(\boldsymbol{\Sigma}_2)) = (1223, 1457) \quad (= (0.172p, 0.204p)), \quad (6.2)$$

where $\hat{\lambda}_{\max}(\boldsymbol{\Sigma}_i)$ is an estimate of the largest eigenvalue due to the noise-reduction methodology by Yata and Aoshima (2013). We concluded that “ $\lambda(\boldsymbol{\Sigma}_i) \in (0, \infty)$ as $p \rightarrow \infty$ ” does not hold and $\boldsymbol{\Sigma}_i$ s are non-sparse because $\lambda_{\max}(\boldsymbol{\Sigma}_i)$ s are very large. Therefore, we do not recommend to apply the classifier by sparse estimation of $\boldsymbol{\Sigma}_i^{-1}$, such as $W_i(\{T_{\tau_{n_i}}(\mathbf{S}_{in_i})\}^{-1})$. Actually, we did not use any classifiers by sparse estimation of $\boldsymbol{\Sigma}_i^{-1}$ in this section. Also, note that the computational cost for the sparse estimation of $\boldsymbol{\Sigma}_i^{-1}$ is very high when p is large.

We constructed the classifiers: DBDA, GQDA, DLDA-bc, DQDA-bc and FS-DQDA, by using the training samples of sizes $n_1 = 27$ and $n_2 = 11$, and checked the accuracy by using the test samples from each π_i . Throughout this section, we set $\gamma = 0.5$ in (5.1) for FS-DQDA. We compared the classifiers with the hard-margin linear support vector machine (HM-LSVM). See Vapnic (1999) for the details. Note that the data sets are linearly separable by a hyperplane because $p > n_1 + n_2$. We emphasize that the computational cost of DBDA, GQDA, DLDA-bc, DQDA-bc or FS-DQDA is as low as HM-LSVM even when $p \geq 10,000$. We summarized misclassification rates in the first block of Table 1. We note that $n_{\min} = 11$ and $n_{\min}^{-1} \log p = 0.81$, so that “ $n_{\min}^{-1} \log p = o(1)$ ” does not hold. That is probably the reason why DLDA-bc, DQDA-bc and FS-DQDA seem to lose the consistency property. See Sections 4 and 5 for the details. On the other hand, DBDA and GQDA gave reasonable performances even when n_i s are small and seem to hold the consistency property. We calculated $\text{tr}(\mathbf{S}_{1 n_1})/\text{tr}(\mathbf{S}_{2 n_2}) = 0.989$ and $(\hat{\Delta}_{i(II)} \text{tr}(\mathbf{S}_{i' n_{i'}})/p)/\hat{\Delta}_{(I)} \approx 1$ for $i \neq i'$. The difference of the trace of the covariance matrices is small and this is probably the reason why DBDA gave a preferable performance. See Section 4.2 for the details. In addition, HM-LSVM also gave a preferable performance. See Hall, Marron and Neeman (2005) for the consistency property of HM-LSVM. For this data set, Cai and Liu (2011) summarized misclassification rates for several other classifiers including a sparse linear classifier called LPD. See Table 6 in Cai and Liu (2011) for the performances of the other classifiers. Note that LPD has the Bayes error rates asymptotically under several sparsity

Table 1: Error rates of the classifiers for samples from Golub et al. (1999).

| Classifier | DBDA | GQDA | DLDA-bc | DQDA-bc | FS-DQDA | HM-LSVM |
|--|------|------|---------|---------|---------|---------|
| Test samples (ALL: 20 and AML: 14) | | | | | | |
| Error rate | 1/34 | 1/34 | 5/34 | 2/34 | 3/34 | 1/34 |
| LOOCV of samples (ALL: 47 and AML: 25) | | | | | | |
| Error rate | 3/72 | 6/72 | 11/72 | 1/72 | 0/72 | 2/72 |

Table 2: Estimates of $(\|\boldsymbol{\mu}_{12}\|^2, \|\boldsymbol{\Sigma}_{12}\|_F^2)$ by $(\hat{\Delta}_{(I)}, \hat{\Delta}_{(S)})$ for Armstrong et al. (2002).

| Case | (a) ALL and MLL | (b) ALL and AML | (c) MLL and AML |
|------------------------------------|------------------------------|-----------------------------|------------------------------|
| $\ \boldsymbol{\mu}_{12}\ ^2$ | 4076 (= 0.324p) | 15050 (= 1.2p) | 8546 (= 0.679p) |
| $\ \boldsymbol{\Sigma}_{12}\ _F^2$ | 1.12×10^8 (= 8863p) | 5.49×10^6 (= 436p) | 1.16×10^8 (= 9192p) |

conditions. We observed that DBDA and GQDA gave the same accuracy as LPD. This is probably because the sparsity conditions do not hold for this data set, so that the Bayes error rates are almost 0. However, the computational cost for DBDA and GQDA is much lower than LPD.

Next, by using all the samples (i.e., 72 samples), we checked the accuracy of the classifiers by the leave-one-out cross-validation (LOOCV). We summarized misclassification rates in the second block of Table 1. We note that $n_{\min} = 24$ and $n_{\min}^{-1} \log p = 0.37$ or $n_{\min} = 25$ and $n_{\min}^{-1} \log p = 0.35$ in this case, so that $n_{\min}^{-1} \log p$ is a little small. We observed that DQDA-bc and FS-DQDA give preferable performances. On the other hand, DLDA-bc gave a poor performance because it does not draw information about heteroscedasticity. For other classifiers, Tan et al. (2005) summarized results of the LOOCV for this data set.

Finally, we analyzed gene expression data given by Armstrong et al. (2002) in which the data set consists of 12582 ($= p$) genes and 72 samples. We had 3 classes of leukemia subtypes: acute lymphoblastic leukemia (ALL: 24 samples), mixed-lineage leukemia (MLL: 20 samples), and acute myeloid leukemia (AML: 28 samples). We considered three cases: (a) ALL and MLL, (b) ALL and AML, and (c) MLL and AML. We standardized each sample by $\mathbf{x}_{ik}/\{\sum_{l=1}^3 \text{tr}(\mathbf{S}_{l n_l})/(3p)\}^{1/2}$ for all i, k , as before. Then, we calculated $(\hat{\Delta}_{(I)}, \hat{\Delta}_{(S)})$ for the three cases. We summarized $(\hat{\Delta}_{(I)}, \hat{\Delta}_{(S)})$ s in Table 2. From Table 2, we concluded that $\boldsymbol{\mu}_{12}$ and $\boldsymbol{\Sigma}_{12}$ are non-sparse for (a) to (c). Also, by using $\lambda_{\max}(\boldsymbol{\Sigma}_i)$, we estimated the largest eigenvalues as 1896, 3206 and 2101 for ALL, MLL and AML, respectively. From this observation, we concluded that $\boldsymbol{\Sigma}_i$ s are non-sparse. We estimated $\text{tr}(\boldsymbol{\Sigma}_{\max}^2)/(n_{\min} \Delta_{(I)}^2)$ and $\lambda_{\max}/\Delta_{(I)}$ by $C_1 = \max\{W_{1n_1}, W_{2n_2}\}/(n_{\min} \hat{\Delta}_{(I)}^2)$ and $C_2 = \max\{\hat{\lambda}_{\max}(\boldsymbol{\Sigma}_1), \hat{\lambda}_{\max}(\boldsymbol{\Sigma}_2)\}/\hat{\Delta}_{(I)}$ in (C-i') and (C-ii'). Then, we had (C_1, C_2) as (0.362, 0.787) for (a), (0.001, 0.14) for (b), and (0.082, 0.375) for (c). Note that $\liminf_{p \rightarrow \infty} \Delta_{\min(II)}/\Delta_{(I)} > 0$ and $\liminf_{p \rightarrow \infty} \Delta_{\min(III)}/\Delta_{(I)} > 0$. From these observations, it is likely that the classifiers by (I) to (III) satisfy (C-i') and (C-ii') specially for (b) and hold the consistency property in (2.2) from Proposition 2.1.

Based on all the samples, we checked the accuracy of the classifiers by using the LOOCV for (a) to (c). We checked the accuracy for 3-class classification as well by

Table 3: Error rates of the classifiers for samples from Armstrong et al. (2002).

| Classifier | DBDA | GQDA | DLDA-bc | DQDA-bc | FS-DQDA | HM-LSVM |
|--|------|------|---------|---------|---------|---------|
| LOOCV of samples from (a) ALL: 24 and MLL: 20 | | | | | | |
| Error rate | 1/44 | 2/44 | 6/44 | 1/44 | 0/44 | 0/44 |
| LOOCV of samples from (b) ALL: 24 and AML: 28 | | | | | | |
| Error rate | 1/52 | 1/52 | 1/52 | 0/52 | 0/52 | 0/52 |
| LOOCV of samples from (c) MLL: 20 and AML: 28 | | | | | | |
| Error rate | 4/48 | 4/48 | 1/48 | 3/48 | 3/48 | 3/48 |
| LOOCV of samples from ALL: 24, MLL: 20 and AML: 28 | | | | | | |
| Error rate | 5/72 | 6/72 | 7/72 | 4/72 | 2/72 | 3/72 |

using the multiclass classification rule given in Remark 1. In the 3-class classification, we used $\hat{\theta}_j$ given in Remark 5 for FS-DQDA and used the one-versus-one approach for HM-LSVM. We summarized misclassification rates in Table 3. We observed that FS-DQDA gives excellent performances. HM-LSVM also gave reasonable performances, however, it does not draw information about the difference of the covariance matrices. See Section 2.2 in Aoshima and Yata (2014) for such an example. As for (b), all the classifiers gave preferable performances. This is probably because the classifiers by (I) to (III) satisfy (C-i') and (C-ii') for (b).

7 Concluding remarks

In this paper, we considered high-dimensional quadratic classifiers in non-sparse settings. The classifier based on the Mahalanobis distance does not always give a preferable performance even when $n_{\min} \rightarrow \infty$ and π_i s are assumed Gaussian, having known covariance matrices. See Sections 1 and 3. We emphasize that the quadratic classifiers proposed in this paper draw information about heterogeneity effectively through both the differences of mean vectors and covariance matrices. See Section 3.4 for the details. If the difference is not sufficiently large, we recommend to use the linear classifiers, DBDA and DLDA-bc (or the original DLDA). They are quite flexible about the conditions to claim the consistency property. See Sections 4.2 and 4.3 for the details. We emphasize that DLDA-bc, DQDA-bc and FS-DQDA can hold the consistency property under at least $n_{\min}^{-1} \log p = o(1)$. Thus we do not recommend to use the classifiers when $n_{\min}^{-1} \log p \neq o(1)$. In such cases, one should use DBDA and GQDA because they hold the consistency property even when n_i s are fixed. See Section 4.2 about the choice between DBDA and GQDA. When $n_{\min}^{-1} \log p = o(1)$, we recommend DQDA-bc and FS-DQDA. Especially, FS-DQDA can claim the consistency property even when $n_{\min}/p \rightarrow 0$ and Δ_{\min} is not sufficiently large. See Section 5.1 for the details. For a choice of $\gamma \in (0, 1)$ in (5.1), one may apply cross-validation procedures or simply choose as $\gamma = 0.5$. Actually, FS-DQDA with $\gamma = 0.5$ gave preferable performances throughout our simulations and real data analyses. On the other hand, even when $n_{\min}^{-1} \log p = o(1)$, we do not recommend to use classifiers by the sparse estimation of Σ_i^{-1} unless (1) the eigenvalues are bounded in the sense that $\lambda(\Sigma_i) \in (0, \infty)$ as $p \rightarrow \infty$, and (2) Σ_i s are sparse in the sense that many elements of Σ_i s are very small. We em-

phasize that “ $\lambda_{\max}(\Sigma_i)$ s are bounded” is a strict condition since the eigenvalues should depend on p and it is probable that $\lambda_{ij} \rightarrow \infty$ as $p \rightarrow \infty$ for the first several j s. See Yata and Aoshima (2013) for the details. Also, the computational cost of the classifiers by the sparse estimation is terribly high.

In conclusion, we hope we have given simpler classifiers which will be more effective in the real world analysis of high-dimensional data.

Appendix A

We give proofs of the theorems. For proofs of the corollaries and the propositions, see Appendix B.

Proof of Theorem 2.1. We consider the case when $\mathbf{x}_0 \in \pi_1$. Under (C-i) and (C-ii), it holds that for $i = 1, 2$

$$\begin{aligned} \text{Var}\{(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T \mathbf{A}_i(\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_i)\} &= \text{tr}(\Sigma_i \mathbf{A}_i \Sigma_1 \mathbf{A}_i)/n_i = o(\Delta_1^2) \\ \text{and } \text{Var}\{(\mathbf{x}_0 - \boldsymbol{\mu}_1 - \bar{\mathbf{x}}_{2n_2} + \boldsymbol{\mu}_2)^T \mathbf{A}_2 \boldsymbol{\mu}_{12}\} &= \boldsymbol{\mu}_{12}^T \mathbf{A}_2 (\Sigma_1 + \Sigma_2/n_2) \mathbf{A}_2 \boldsymbol{\mu}_{12} = o(\Delta_1^2) \end{aligned} \quad (\text{A.1})$$

from the fact that

$$\boldsymbol{\mu}_{12}^T \mathbf{A}_2 \Sigma_2 \mathbf{A}_2 \boldsymbol{\mu}_{12} \leq \boldsymbol{\mu}_{12}^T \mathbf{A}_2 \boldsymbol{\mu}_{12} \lambda_{\max}(\mathbf{A}_2^{1/2} \Sigma_2 \mathbf{A}_2^{1/2}) \leq \Delta_1 \text{tr}\{(\Sigma_2 \mathbf{A}_2)^2\}^{1/2} = o(n_2 \Delta_1^2)$$

under (C-i). Note that $(\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_i)^T \mathbf{A}_i(\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_i) - \text{tr}(\mathbf{A}_i \mathbf{S}_{in_i})/n_i = \sum_{k \neq k'}^{n_i} (\mathbf{x}_{ik} - \boldsymbol{\mu}_i)^T \mathbf{A}_i(\mathbf{x}_{ik'} - \boldsymbol{\mu}_i)/\{n_i(n_i - 1)\}$. Then, under (C-i) it follows that for $i = 1, 2$

$$\text{Var}\{(\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_i)^T \mathbf{A}_i(\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_i) - \text{tr}(\mathbf{A}_i \mathbf{S}_{in_i})/n_i\} = O[\text{tr}\{(\Sigma_i \mathbf{A}_i)^2\}/n_i^2] = o(\Delta_1^2). \quad (\text{A.2})$$

Then, by using Chebyshev’s inequality, from (A.1) and (A.2), we find that

$$W_2(\mathbf{A}_2) - W_1(\mathbf{A}_1) = \text{tr}[\{(\mathbf{x}_0 - \boldsymbol{\mu}_1)(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T - \Sigma_1\}(\mathbf{A}_2 - \mathbf{A}_1)] + \Delta_1 + o_P(\Delta_1). \quad (\text{A.3})$$

Here, under (A-i) and (C-iii), it follows that

$$\text{Var}(\text{tr}[\{(\mathbf{x}_0 - \boldsymbol{\mu}_1)(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T - \Sigma_1\}(\mathbf{A}_2 - \mathbf{A}_1)]) = O(\text{tr}[\{\Sigma_1(\mathbf{A}_2 - \mathbf{A}_1)\}^2]) = o(\Delta_1^2). \quad (\text{A.4})$$

Thus by combining (A.3) with (A.4), under (A-i) and (C-i) to (C-iii), we obtain that $\{W_2(\mathbf{A}_2) - W_1(\mathbf{A}_1)\}/\Delta_1 = 1 + o_P(1)$, so that $P\{W_2(\mathbf{A}_2) - W_1(\mathbf{A}_1) > 0\} \rightarrow 1$. When $\mathbf{x}_0 \in \pi_2$, we have the same arguments. The proof is completed. \square

Proof of Theorem 3.1. Note that $\text{tr}\{(\Sigma_i \mathbf{A}_i)^2\}/n_i^2 = o(\delta_i^2)$, $i = 1, 2$. Then, similar to (A.1) to (A.4), under (A-i) and (C-iv) to (C-vi), we have that as $m \rightarrow \infty$

$$W_{i'}(\mathbf{A}_{i'}) - W_i(\mathbf{A}_i) - \Delta_i = 2(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T \{\mathbf{A}_i(\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_i) - \mathbf{A}_{i'}(\bar{\mathbf{x}}_{i'n_{i'}} - \boldsymbol{\mu}_{i'})\} + o_P(\delta_i) \quad (\text{A.5})$$

when $\mathbf{x}_0 \in \pi_i$ ($i' \neq i$). Note that $2\omega_i/\delta_i \rightarrow 1$ as $m \rightarrow \infty$ for $i = 1, 2$, under (C-vi), where $\omega_i = \{\text{tr}\{(\Sigma_i \mathbf{A}_i)^2\}/n_i + \text{tr}(\Sigma_i \mathbf{A}_{i'} \Sigma_{i'} \mathbf{A}_{i'})/n_{i'}\}^{1/2}$ ($i' \neq i$) in view of Lemma B.1 of Appendix B. Then, by combining Lemma B.1 with (A.5), we conclude the results. \square

Proof of Theorem 3.2. Similar to (A.5), under (A-i), (C-iv) and (C-v), we have that as $m \rightarrow \infty$

$$\begin{aligned} W_{i'}(\mathbf{A}_{i'}) - W_i(\mathbf{A}_i) - \Delta_i = & 2(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T \{ \mathbf{A}_i(\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_i) \\ & - \mathbf{A}_{i'}(\bar{\mathbf{x}}_{i'n_{i'}} - \boldsymbol{\mu}_{i'} + (-1)^i \boldsymbol{\mu}_{12}) \} + o_P(\delta_i) \end{aligned} \quad (\text{A.6})$$

when $\mathbf{x}_0 \in \pi_i$ ($i' \neq i$). Then, by combining Lemma B.2 of Appendix B with (A.6), we conclude the results. \square

Proof of Theorem 5.1. By using (B.23) and (B.24) in Appendix B, we claim the result. \square

Appendix B

Throughout, we consider the eigen-decomposition of \mathbf{A}_i by $\mathbf{A}_i = \mathbf{H}_{i(A)} \boldsymbol{\Lambda}_{i(A)} \mathbf{H}_{i(A)}^T$ for $i = 1, 2$, where $\boldsymbol{\Lambda}_{i(A)} = \text{diag}(\lambda_{i1(A)}, \dots, \lambda_{ip(A)})$ having eigenvalues such as $\lambda_{i1(A)} \geq \dots \geq \lambda_{ip(A)} > 0$ and $\mathbf{H}_{i(A)} = [\mathbf{h}_{i1(A)}, \dots, \mathbf{h}_{ip(A)}]$ is an orthogonal matrix of the corresponding eigenvectors. Let $a_{i(j)}$ be the j -th diagonal element of \mathbf{A}_i for $j = 1, \dots, p$ ($i = 1, 2$). Let $\tilde{\mathbf{x}}_{1k} = \mathbf{A}_1^{1/2}(\mathbf{x}_{1k} - \boldsymbol{\mu}_1)$ and $\tilde{\mathbf{x}}_{2k} = \mathbf{A}_2^{-1/2} \mathbf{A}_2(\mathbf{x}_{2k} - \boldsymbol{\mu}_2)$ for $k = 1, \dots, n_i$. Let $\tilde{\boldsymbol{\Sigma}}_1 = \mathbf{A}_1^{1/2} \boldsymbol{\Sigma}_1 \mathbf{A}_1^{1/2}$, $\tilde{\boldsymbol{\Sigma}}_2 = \mathbf{A}_1^{-1/2} \mathbf{A}_2 \boldsymbol{\Sigma}_2 \mathbf{A}_2 \mathbf{A}_1^{-1/2}$, $\tilde{\boldsymbol{\Gamma}}_1 = [\tilde{\gamma}_{11}, \dots, \tilde{\gamma}_{1q_1}] = \mathbf{A}_1^{1/2} \boldsymbol{\Gamma}_1$ and $\tilde{\boldsymbol{\Gamma}}_2 = [\tilde{\gamma}_{21}, \dots, \tilde{\gamma}_{2q_2}] = \mathbf{A}_1^{-1/2} \mathbf{A}_2 \boldsymbol{\Gamma}_2$. Note that $\text{Var}(\tilde{\mathbf{x}}_{ij}) = \tilde{\boldsymbol{\Gamma}}_i \tilde{\boldsymbol{\Gamma}}_i^T = \sum_{j=1}^{q_i} \tilde{\gamma}_{ij} \tilde{\gamma}_{ij}^T = \tilde{\boldsymbol{\Sigma}}_i$, $i = 1, 2$. Let $\hat{\mathbf{B}}_i = \hat{\mathbf{A}}_i - \mathbf{A}_i$ for $i = 1, 2$. Let $x_{oijk} = x_{ijk} - \mu_{ij}$ for $j = 1, \dots, p$ ($i = 1, 2$; $k = 1, \dots, n_i$).

Proof of Proposition 1.1. We can write that $\text{tr}(\mathbf{A}_i^{-1} \mathbf{A}_{i'}) = \sum_{j=1}^p \mathbf{h}_{ij(A)}^T \mathbf{A}_{i'} \mathbf{h}_{ij(A)} / \lambda_{ij(A)}$. Note that $\sum_{j=1}^p \mathbf{h}_{ij(A)}^T \mathbf{A}_{i'} \mathbf{h}_{ij(A)} = \text{tr}(\mathbf{A}_{i'})$ and $\sum_{j=1}^t (\mathbf{h}_{ij(A)}^T \mathbf{A}_{i'} \mathbf{h}_{ij(A)} - \lambda_{i'j(A)}) \leq 0$ for any $t \in \{1, \dots, p\}$. Then, by noting that $\lambda_{i1(A)} \geq \dots \geq \lambda_{ip(A)} > 0$, we have that

$$\begin{aligned} \text{tr}(\mathbf{A}_i^{-1} \mathbf{A}_{i'}) &= \frac{\lambda_{i'1(A)}}{\lambda_{i1(A)}} + \frac{\mathbf{h}_{i1(A)}^T \mathbf{A}_{i'} \mathbf{h}_{i1(A)} - \lambda_{i'1(A)}}{\lambda_{i1(A)}} + \sum_{j=2}^p \frac{\mathbf{h}_{ij(A)}^T \mathbf{A}_{i'} \mathbf{h}_{ij(A)}}{\lambda_{ij(A)}} \\ &\geq \sum_{j=1}^2 \frac{\lambda_{i'j(A)}}{\lambda_{ij(A)}} + \sum_{j=1}^2 \frac{\mathbf{h}_{ij(A)}^T \mathbf{A}_{i'} \mathbf{h}_{ij(A)} - \lambda_{i'j(A)}}{\lambda_{i2(A)}} + \sum_{j=3}^p \frac{\mathbf{h}_{ij(A)}^T \mathbf{A}_{i'} \mathbf{h}_{ij(A)}}{\lambda_{ij(A)}} \\ &\vdots \\ &\geq \sum_{j=1}^p \frac{\lambda_{i'j(A)}}{\lambda_{ij(A)}} + \sum_{j=1}^p \frac{\mathbf{h}_{ij(A)}^T \mathbf{A}_{i'} \mathbf{h}_{ij(A)} - \lambda_{i'j(A)}}{\lambda_{ip(A)}} = \sum_{j=1}^p \frac{\lambda_{i'j(A)}}{\lambda_{ij(A)}}. \end{aligned} \quad (\text{B.7})$$

Thus, when $\text{tr}\{\boldsymbol{\Sigma}_i(\mathbf{A}_{i'} - \mathbf{A}_i)\} = \text{tr}(\mathbf{A}_i^{-1} \mathbf{A}_{i'}) - p$, it holds that

$$\Delta_i \geq \sum_{j=1}^p \{ \lambda_{i'j(A)} / \lambda_{ij(A)} - 1 + \log(\lambda_{ij(A)} / \lambda_{i'j(A)}) \} \geq 0$$

from the fact that $c - 1 + \log c^{-1} \geq 0$ for any positive constant c . Note that $\lambda_{1j(A)} \neq \lambda_{2j(A)}$ or $\mathbf{h}_{ij(A)}^T \mathbf{A}_{i'} \mathbf{h}_{ij(A)} < \lambda_{i'j(A)}$ for some j when $\mathbf{A}_1 \neq \mathbf{A}_2$. Since $c - 1 + \log c^{-1} > 0$ when $c \neq 1$, it holds that $\Delta_i > 0$ when $\lambda_{1j(A)} \neq \lambda_{2j(A)}$ for some j . From (B.7), if

$\mathbf{h}_{ij(A)}^T \mathbf{A}_{i'} \mathbf{h}_{ij(A)} < \lambda_{i'j(A)}$ for some j , it follows that $\text{tr}(\mathbf{A}_i^{-1} \mathbf{A}_{i'}) > \sum_{j=1}^p (\lambda_{i'j(A)} / \lambda_{ij(A)})$, so that $\Delta_i > 0$. When $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, it holds that $\Delta_i \geq \boldsymbol{\mu}_{12}^T \mathbf{A}_{i'} \boldsymbol{\mu}_{12} > 0$. Hence, it concludes the results. \square

Proof of Proposition 2.1. We note that

$$\begin{aligned} \Delta_{iA} &\leq \boldsymbol{\mu}_{12}^T \mathbf{A}_{i'} \boldsymbol{\mu}_{12} \lambda_{\max}(\mathbf{A}_{i'}^{1/2} \boldsymbol{\Sigma}_i \mathbf{A}_{i'}^{1/2}) \leq \Delta_i \lambda_{\max}(\mathbf{A}_{i'}^{1/2} \boldsymbol{\Sigma}_i \mathbf{A}_{i'}^{1/2}) \\ \text{and } \text{tr}(\boldsymbol{\Sigma}_i \mathbf{A}_{i'} \boldsymbol{\Sigma}_{i'} \mathbf{A}_{i'}) &\leq \text{tr}\{(\boldsymbol{\Sigma}_i \mathbf{A}_{i'})^2\}^{1/2} \text{tr}\{(\boldsymbol{\Sigma}_{i'} \mathbf{A}_{i'})^2\}^{1/2}. \end{aligned} \quad (\text{B.8})$$

When $\limsup_{p \rightarrow \infty} \lambda_{i1(A)} < \infty$, $i = 1, 2$, it holds that

$$\begin{aligned} \lambda_{\max}(\mathbf{A}_{i'}^{1/2} \boldsymbol{\Sigma}_i \mathbf{A}_{i'}^{1/2}) &\leq \lambda_{i1} \lambda_{\max}(\mathbf{A}_{i'}) = \lambda_{i1} \lambda_{i'1(A)} = O(\lambda_{i1}) \quad \text{and} \\ \text{tr}\{(\boldsymbol{\Sigma}_l \mathbf{A}_{l'})^2\} &\leq \text{tr}(\boldsymbol{\Sigma}_l \mathbf{A}_{l'} \boldsymbol{\Sigma}_l) \lambda_{l'1(A)} \leq \text{tr}(\boldsymbol{\Sigma}_l^2) \lambda_{l'1(A)}^2 = O\{\text{tr}(\boldsymbol{\Sigma}_l^2)\} \end{aligned} \quad (\text{B.9})$$

for all l, l' . By combining (B.8) with (B.9), (C-i') and (C-ii') imply (C-i) and (C-ii).

Next, for (C-iii), it holds that $\text{tr}[\{\boldsymbol{\Sigma}_i(\mathbf{A}_1 - \mathbf{A}_2)\}^2] \leq \lambda_{i1} \text{tr}\{(\mathbf{A}_1 - \mathbf{A}_2) \boldsymbol{\Sigma}_i (\mathbf{A}_1 - \mathbf{A}_2)\}$. When \mathbf{A}_i s are diagonal matrices such as $\mathbf{A}_i = \text{diag}(a_{i(1)}, \dots, a_{i(p)})$, $i = 1, 2$, it holds that $\Delta_i \geq \sum_{j=1}^p \{a_{i'(j)} / a_{i(j)} - 1 - \log(a_{i'(j)} / a_{i(j)})\}$ and $\text{tr}\{(\mathbf{A}_1 - \mathbf{A}_2) \boldsymbol{\Sigma}_i (\mathbf{A}_1 - \mathbf{A}_2)\} = \sum_{j=1}^p \sigma_{i(j)} (a_{1(j)} - a_{2(j)})^2$. Note that $a_{i(j)} \in (0, \infty)$ as $p \rightarrow \infty$ for all i, j , under $\lambda(\mathbf{A}_i) \in (0, \infty)$ as $p \rightarrow \infty$ for $i = 1, 2$. By Taylor expansion, we claim that

$$a_{i'(j)} / a_{i(j)} - 1 - \log(a_{i'(j)} / a_{i(j)}) \geq a_{i(j)}^{-2} (a_{1(j)} - a_{2(j)})^2 / (2 \max\{1, a_{i'(j)}^2 / a_{i(j)}^2\}).$$

Then, it follows that $\sum_{j=1}^p \sigma_{i(j)} (a_{1(j)} - a_{2(j)})^2 = O(\Delta_i)$ because $\sigma_{i(j)} \in (0, \infty)$ as $p \rightarrow \infty$ for all i, j . Thus we have that $\text{tr}[\{\boldsymbol{\Sigma}_i(\mathbf{A}_1 - \mathbf{A}_2)\}^2] = O(\Delta_i \lambda_{i1})$. It concludes the results. \square

Proofs of Corollaries 2.1 and 2.2. From Theorem 2.1 and Proposition 2.1, we can claim Corollaries 2.1 and 2.2 straightforwardly. \square

Proof of Proposition 2.2. We first consider the case when $\liminf_{p \rightarrow \infty} \sum_{j=1}^p |\lambda_{ij} / \lambda_{i'j} - 1| / p > 0$. When $c_{1j} < |\lambda_{ij} / \lambda_{i'j} - 1| < c_{2j}$ for some constants $c_{1j} (> 0)$ and $c_{2j} (< \infty)$, by Taylor expansion, it holds that

$$\lambda_{ij} / \lambda_{i'j} - 1 - \log(\lambda_{ij} / \lambda_{i'j}) \geq \frac{(\lambda_{ij} / \lambda_{i'j} - 1)^2}{2 \max\{1, \lambda_{ij}^2 / \lambda_{i'j}^2\}} \geq \frac{c_{1j} |\lambda_{ij} / \lambda_{i'j} - 1|}{2(c_{2j} + 1)^2}.$$

When $\lambda_{ij} / \lambda_{i'j} \rightarrow \infty$ as $p \rightarrow \infty$, it holds that for sufficiently large p

$$\lambda_{ij} / \lambda_{i'j} - 1 - \log(\lambda_{ij} / \lambda_{i'j}) > |\lambda_{ij} / \lambda_{i'j} - 1| / 2.$$

Thus, when $\liminf_{p \rightarrow \infty} \sum_{j=1}^p |\lambda_{ij} / \lambda_{i'j} - 1| / p > 0$, it follows that $\liminf_{p \rightarrow \infty} \Delta_{i(IV)} / p \geq \liminf_{p \rightarrow \infty} \sum_{j=1}^p \{\lambda_{ij} / \lambda_{i'j} - 1 - \log(\lambda_{ij} / \lambda_{i'j})\} / p > 0$ from (B.7).

Next, we consider the case when $\liminf_{p \rightarrow \infty} |\text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'}^{-1}) / p - 1| > 0$. We note that $\text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'}^{-1}) \geq \sum_{j=1}^p \lambda_{ij} / \lambda_{i'j}$ from (B.7). When $\text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'}^{-1}) / (\sum_{j=1}^p \lambda_{ij} / \lambda_{i'j}) \rightarrow 1$ as $p \rightarrow \infty$, it holds that $\liminf_{p \rightarrow \infty} |\sum_{j=1}^p (\lambda_{ij} / \lambda_{i'j}) / p - 1| > 0$ under $\liminf_{p \rightarrow \infty} |\text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'}^{-1}) / p - 1| > 0$

0. It follows that $\liminf_{p \rightarrow \infty} \Delta_{i(IV)}/p > 0$ from the fact that $\sum_{j=1}^p |\lambda_{ij}/\lambda_{i'j} - 1|/p \geq |\sum_{j=1}^p (\lambda_{ij}/\lambda_{i'j})/p - 1|$. On the other hand, we note that

$$\Delta_{i(IV)} \geq \text{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1}) - p - \sum_{j=1}^p \log(\lambda_{ij}/\lambda_{i'j}) \geq \text{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1}) - \sum_{j=1}^p (\lambda_{ij}/\lambda_{i'j})$$

because $\sum_{j=1}^p \{\lambda_{ij}/\lambda_{i'j} - 1 - \log(\lambda_{ij}/\lambda_{i'j})\} \geq 0$. Thus, when $\sum_{j=1}^p (\lambda_{ij}/\lambda_{i'j})/p - 1 \rightarrow 0$ as $p \rightarrow \infty$ and $\liminf_{p \rightarrow \infty} \{\text{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_{i'}^{-1})/(\sum_{j=1}^p \lambda_{ij}/\lambda_{i'j})\} > 1$, we have that $\liminf_{p \rightarrow \infty} \Delta_{i(IV)}/p > 0$. Hence, it concludes the results. \square

Proof of Proposition 3.1. Under $\limsup_{p \rightarrow \infty} \lambda_{i1(A)} < \infty$ for $i = 1, 2$, we have that $\text{tr}\{(\mathbf{\Sigma}_{i'} \mathbf{A}_{i'})^2\} = O\{\text{tr}(\mathbf{\Sigma}_{i'}^2)\}$ and

$$\begin{aligned} \text{tr}\{(\mathbf{\Sigma}_i \mathbf{A}_l \mathbf{\Sigma}_l \mathbf{A}_l)^2\} &= \text{tr}\{(\mathbf{\Sigma}_i^{1/2} \mathbf{A}_l \mathbf{\Sigma}_l \mathbf{A}_l \mathbf{\Sigma}_i^{1/2})^2\} \\ &\leq \lambda_{\max}(\mathbf{\Sigma}_i^{1/2} \mathbf{A}_l \mathbf{\Sigma}_l \mathbf{A}_l \mathbf{\Sigma}_i^{1/2}) \text{tr}(\mathbf{\Sigma}_i^{1/2} \mathbf{A}_l \mathbf{\Sigma}_l \mathbf{A}_l \mathbf{\Sigma}_i^{1/2}) \\ &\leq \lambda_{\max}(\mathbf{\Sigma}_i^{1/2} \mathbf{A}_l^2 \mathbf{\Sigma}_i^{1/2}) \lambda_{l1} \delta_i^2 n_l = O(\lambda_{i1} \lambda_{l1} \delta_i^2 n_l); \text{ and} \\ \boldsymbol{\mu}_{12}^T \mathbf{A}_{i'} \mathbf{\Sigma}_l \mathbf{A}_{i'} \boldsymbol{\mu}_{12} &\leq \|\boldsymbol{\mu}_{12}\|^2 \lambda_{\max}(\mathbf{A}_{i'} \mathbf{\Sigma}_l \mathbf{A}_{i'}) = O(\|\boldsymbol{\mu}_{12}\|^2 \lambda_{l1}) \text{ for } l = i, i'. \end{aligned}$$

Then, when $\limsup_{p \rightarrow \infty} \lambda_{i1(A)} < \infty$, $i = 1, 2$, (C-iv') and (C-vi') imply (C-iv) and (C-vi), respectively. Similar to Proof of Proposition 2.1, we can claim the result for (C-v') from $\text{tr}\{(\mathbf{A}_1 - \mathbf{A}_2)^2\} = \sum_{j=1}^p (a_{1(j)} - a_{2(j)})^2$ when \mathbf{A}_i s are diagonal matrices. The proof is completed. \square

Lemma B.1. Let $\omega_i = \{\text{tr}\{(\mathbf{\Sigma}_i \mathbf{A}_i)^2\}/n_i + \text{tr}(\mathbf{\Sigma}_i \mathbf{A}_{i'} \mathbf{\Sigma}_{i'} \mathbf{A}_{i'})/n_{i'}\}^{1/2}$ for $i = 1, 2$ ($i' \neq i$). Then, under (A-i), (C-iv) and (C-vi), we have that

$$(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T \{\mathbf{A}_i(\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_i) - \mathbf{A}_{i'}(\bar{\mathbf{x}}_{i'n_{i'}} - \boldsymbol{\mu}_{i'})\}/\omega_i \Rightarrow N(0, 1) \text{ as } m \rightarrow \infty$$

when $\mathbf{x}_0 \in \pi_i$ for $i = 1, 2$ ($i' \neq i$).

Proof of Lemma B.1. We consider the case when $i = 1$ ($i' = 2$) and $\mathbf{x}_0 \in \pi_1$. Let $\tilde{\mathbf{x}}_0 = \mathbf{A}_1^{1/2}(\mathbf{x}_0 - \boldsymbol{\mu}_1)$. Then, it holds that $\text{Var}(\tilde{\mathbf{x}}_0 | \mathbf{x}_0 \in \pi_1) = \text{Var}(\tilde{\mathbf{x}}_{1k}) = \tilde{\mathbf{\Sigma}}_1$. Let

$$v_k = \tilde{\mathbf{x}}_0^T \tilde{\mathbf{x}}_{1k}/(n_1 \omega_1), \quad k = 1, \dots, n_1, \text{ and } v_{n_1+k} = -\tilde{\mathbf{x}}_0^T \tilde{\mathbf{x}}_{2k}/(n_2 \omega_1), \quad k = 1, \dots, n_2.$$

Note that $\sum_{k=1}^{n_1+n_2} E(v_k^2) = 1$ and

$$\sum_{k=1}^{n_1+n_2} v_k = (\mathbf{x}_0 - \boldsymbol{\mu}_1)^T \{\mathbf{A}_1(\bar{\mathbf{x}}_{1n_1} - \boldsymbol{\mu}_1) - \mathbf{A}_2(\bar{\mathbf{x}}_{2n_2} - \boldsymbol{\mu}_2)\}/\omega_1.$$

Then, it holds that $E(v_k | v_{k-1}, \dots, v_1) = 0$ for $k = 2, \dots, n_1 + n_2$. We consider applying the martingale central limit theorem given by McLeish (1974). Under (A-i), we can write that $\tilde{\mathbf{x}}_{1l} = \tilde{\mathbf{\Gamma}}_1 \mathbf{y}_{1l}$ and $\tilde{\mathbf{x}}_{2l} = \tilde{\mathbf{\Gamma}}_2 \mathbf{y}_{2l}$. Then, in a way similar to the equations (23) and (24) in Aoshima and Yata (2014), we can evaluate that under (A-i)

$$(n_s \omega_1)^4 E(v_k^4) = 3\text{tr}(\tilde{\mathbf{\Sigma}}_1 \tilde{\mathbf{\Sigma}}_s)^2 + O[\text{tr}\{(\tilde{\mathbf{\Sigma}}_1 \tilde{\mathbf{\Sigma}}_s)^2\}] \quad \text{and} \quad (\text{B.10})$$

$$\begin{aligned} (n_s n_{s'})^2 \omega_1^4 E(v_k^2 v_{k'}^2) &= \text{tr}(\tilde{\mathbf{\Sigma}}_1 \tilde{\mathbf{\Sigma}}_s) \text{tr}(\tilde{\mathbf{\Sigma}}_1 \tilde{\mathbf{\Sigma}}_{s'}) + 2\text{tr}(\tilde{\mathbf{\Sigma}}_1 \tilde{\mathbf{\Sigma}}_s \tilde{\mathbf{\Sigma}}_1 \tilde{\mathbf{\Sigma}}_{s'}) \\ &\quad + O[\{\text{tr}(\tilde{\mathbf{\Sigma}}_1 \tilde{\mathbf{\Sigma}}_s \tilde{\mathbf{\Sigma}}_1 \tilde{\mathbf{\Sigma}}_s) \text{tr}(\tilde{\mathbf{\Sigma}}_1 \tilde{\mathbf{\Sigma}}_{s'} \tilde{\mathbf{\Sigma}}_1 \tilde{\mathbf{\Sigma}}_{s'})\}^{1/2}] \end{aligned} \quad (\text{B.11})$$

for $k \neq k'$, where $s = 1$ for $k \in [1, \dots, n_1]$, $s = 2$ for $k \in [n_1 + 1, \dots, n_1 + n_2]$, $s' = 1$ for $k' \in [1, \dots, n_1]$, and $s' = 2$ for $k' \in [n_1 + 1, \dots, n_1 + n_2]$. Note that $\text{tr}(\tilde{\Sigma}_1^4) \leq \text{tr}(\tilde{\Sigma}_1^2)^2$ and $\text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2\} \leq \text{tr}(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2$. Then, by using Chebyshev's inequality and Schwarz's inequality, from (B.10), under (A-i), we have that for Lindeberg's condition

$$\sum_{k=1}^{n_1+n_2} E\{v_k^2 I(v_k^2 \geq \tau)\} \leq \sum_{k=1}^{n_1+n_2} \frac{E(v_k^4)}{\tau} = O\left[\frac{\text{tr}(\tilde{\Sigma}_1^2)^2/n_1^3 + \text{tr}(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2/n_2^3}{\omega_1^4}\right] \rightarrow 0$$

as $m \rightarrow \infty$ for any $\tau > 0$, where $I(\cdot)$ is the indicator function. Since $2\omega_1/\delta_1 = 1 + o(1)$ under (C-vi), we note that

$$\begin{aligned} \frac{\text{tr}(\tilde{\Sigma}_1^4)}{n_1^2 \omega_1^4} &\rightarrow 0, \quad \frac{\text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2\}}{n_2^2 \omega_1^4} = \frac{\text{tr}\{(\Sigma_1 \mathbf{A}_2 \Sigma_2 \mathbf{A}_2)^2\}}{n_2^2 \omega_1^4} \rightarrow 0, \\ \text{and } \frac{\text{tr}(\tilde{\Sigma}_1^3 \tilde{\Sigma}_2)}{n_1 n_2 \omega_1^4} &\leq \frac{\text{tr}(\tilde{\Sigma}_1^4)^{1/2} \text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2\}^{1/2}}{n_1 n_2 \omega_1^4} \rightarrow 0 \end{aligned}$$

under (C-iv). Then, by using Chebyshev's inequality, from (B.10) and (B.11), under (A-i), (C-iv) and (C-vi), we have that for any $\tau > 0$

$$\begin{aligned} P\left(\left|\sum_{k=1}^{n_1+n_2} v_k^2 - 1\right| \geq \tau\right) \\ = O\left[\frac{\text{tr}(\tilde{\Sigma}_1^4)/n_1^2 + \text{tr}(\tilde{\Sigma}_1^4)^{1/2} \text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2\}^{1/2}/(n_1 n_2) + \text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_2)^2\}/n_2^2}{\omega_1^4}\right] + o(1) \rightarrow 0 \end{aligned}$$

as $m \rightarrow \infty$, so that $\sum_{k=1}^{n_1+n_2} v_k^2 = 1 + o_P(1)$. Hence, by using the martingale central limit theorem, we obtain that $\sum_{k=1}^{n_1+n_2} v_k \Rightarrow N(0, 1)$ as $m \rightarrow \infty$ under (A-i), (C-iv) and (C-vi). Hence, we conclude the result when $i = 1$. For the case when $i = 2$, we can have the same arguments. The proof is completed. \square

Lemma B.2. *Under (A-ii), (C-iv) and (C-vii), we have that*

$$2(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T \{\mathbf{A}_i(\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_i) - \mathbf{A}_{i'}(\bar{\mathbf{x}}_{i'n_{i'}} - \boldsymbol{\mu}_{i'} + (-1)^i \boldsymbol{\mu}_{12})\} / \delta_i \Rightarrow N(0, 1)$$

as $m \rightarrow \infty$ when $\mathbf{x}_0 \in \pi_i$ for $i = 1, 2$ ($i' \neq i$).

Proof of Lemma B.2. We consider the case when $i = 1$ ($i' = 2$) and $\mathbf{x}_0 \in \pi_1$. Let $\mathbf{x}_0 - \boldsymbol{\mu}_1 = \boldsymbol{\Gamma}_1 \mathbf{y}_0$ and $\mathbf{y}_0 = (y_{01}, \dots, y_{0q_1})^T$. Under (A-ii), y_{0s} , $s = 1, \dots, q_1$, are independent. Let $\dot{\mathbf{x}}_{ln_l} = \sum_{k=1}^{n_l} \tilde{\mathbf{x}}_{lk}/n_l$, $l = 1, 2$, $\tilde{\boldsymbol{\mu}} = \mathbf{A}_1^{-1/2} \mathbf{A}_2 \boldsymbol{\mu}_{12}$ and

$$w_s = 2y_{0j} \tilde{\gamma}_{1s}^T (\dot{\mathbf{x}}_{1n_1} - \dot{\mathbf{x}}_{2n_2} + \tilde{\boldsymbol{\mu}}) / \delta_1, \quad s = 1, \dots, q_1.$$

Note that $q_1 \geq p$, $E(w_s) = 0$, $s = 1, \dots, q_1$, $\sum_{s=1}^{q_1} E(w_s^2) = 1$ and

$$\sum_{s=1}^{q_1} w_s = 2(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T \{\mathbf{A}_1(\bar{\mathbf{x}}_{1n_1} - \boldsymbol{\mu}_1) - \mathbf{A}_2(\bar{\mathbf{x}}_{2n_2} - \boldsymbol{\mu}_2 - \boldsymbol{\mu}_{12})\} / \delta_1.$$

Also, note that $E(w_s|w_{s-1}, \dots, w_1) = 0$ for $s = 2, \dots, q_1$, under (A-ii). We consider applying the martingale central limit theorem. Let $M_{ls} = E(y_{lsk}^3)$ for all l, s . Note that $\limsup_{p \rightarrow \infty} |M_{ls}| < \infty$ for all l, s , under (A-ii) because $\limsup_{p \rightarrow \infty} E(y_{lsk}^4) < \infty$. Then, by using Schwarz's inequality and the arithmetic mean-geometric mean inequality, we can evaluate that under (A-ii)

$$\begin{aligned}
E\{(\tilde{\gamma}_{1s}^T \dot{\mathbf{x}}_{ln_l})^2 (\tilde{\gamma}_{1t}^T \dot{\mathbf{x}}_{ln_l})^2\} &= \{1 + o(1)\} \tilde{\gamma}_{1s}^T \tilde{\Sigma}_l \tilde{\gamma}_{1s} \tilde{\gamma}_{1t}^T \tilde{\Sigma}_l \tilde{\gamma}_{1t} / n_l^2 + O\{(\tilde{\gamma}_{1s}^T \tilde{\Sigma}_l \tilde{\gamma}_{1t} / n_l)^2\}; \text{ and} \\
|E\{(\tilde{\gamma}_{1s}^T \dot{\mathbf{x}}_{ln_l})^2 \tilde{\gamma}_{1t}^T \dot{\mathbf{x}}_{ln_l} \tilde{\gamma}_{1t}^T \tilde{\mu}\}| &= \left| \sum_{u=1}^{q_l} (\tilde{\gamma}_{1s}^T \tilde{\gamma}_{lu})^2 \tilde{\gamma}_{1t}^T \tilde{\gamma}_{lu} \tilde{\gamma}_{1t}^T \tilde{\mu} M_{lu} / n_l^2 \right| \\
&\leq \{E(\tilde{\gamma}_{1s}^T \dot{\mathbf{x}}_{ln_l})^4\}^{1/2} \{E(\tilde{\gamma}_{1t}^T \dot{\mathbf{x}}_{ln_l} \tilde{\gamma}_{1t}^T \tilde{\mu})^2\}^{1/2} \\
&= O\{\tilde{\gamma}_{1s}^T \tilde{\Sigma}_l \tilde{\gamma}_{1s} (\tilde{\gamma}_{1t}^T \tilde{\Sigma}_l \tilde{\gamma}_{1t} / n_l)^{1/2} |\tilde{\gamma}_{1t}^T \tilde{\mu}| / n_l\} \\
&= O[\tilde{\gamma}_{1s}^T \tilde{\Sigma}_l \tilde{\gamma}_{1s} \{\tilde{\gamma}_{1t}^T \tilde{\Sigma}_l \tilde{\gamma}_{1t} / n_l + (\tilde{\gamma}_{1t}^T \tilde{\mu})^2\} / n_l] \\
&= O[\{\tilde{\gamma}_{1s}^T \tilde{\Sigma}_l \tilde{\gamma}_{1s} / n_l\}^2 + \{\tilde{\gamma}_{1t}^T \tilde{\Sigma}_l \tilde{\gamma}_{1t} / n_l\}^2 + (\tilde{\gamma}_{1t}^T \tilde{\mu})^4], \quad l = 1, 2
\end{aligned}$$

for all s, t . Then, we have that for all s, t

$$\delta_1^4 E(w_s^4) = O\left[\sum_{l=1}^2 \{\tilde{\gamma}_{1s}^T \tilde{\Sigma}_l \tilde{\gamma}_{1s} / n_l\}^2 + (\tilde{\gamma}_{1s}^T \tilde{\mu})^4\right] \quad \text{and} \quad (\text{B.12})$$

$$\begin{aligned}
(\delta_1/2)^4 \frac{E(w_s^2 w_t^2)}{E(y_{0s}^2 y_{0t}^2)} - \tilde{\gamma}_{1s}^T \left(\sum_{l=1}^2 \tilde{\Sigma}_l / n_l + \tilde{\mu} \tilde{\mu}^T \right) \tilde{\gamma}_{1s} \tilde{\gamma}_{1t}^T \left(\sum_{l=1}^2 \tilde{\Sigma}_l / n_l + \tilde{\mu} \tilde{\mu}^T \right) \tilde{\gamma}_{1t} \\
= 2 \sum_{l=1}^2 (-1)^{l+1} \sum_{u=1}^{q_l} \{(\tilde{\gamma}_{1s}^T \tilde{\gamma}_{lu})^2 \tilde{\gamma}_{1t}^T \tilde{\gamma}_{lu} \tilde{\gamma}_{1t}^T + (\tilde{\gamma}_{1t}^T \tilde{\gamma}_{lu})^2 \tilde{\gamma}_{1s}^T \tilde{\gamma}_{lu} \tilde{\gamma}_{1s}^T\} \tilde{\mu} M_{lu} / n_l^2 \\
+ o\left[\sum_{l=1}^2 \tilde{\gamma}_{1s}^T \tilde{\Sigma}_l \tilde{\gamma}_{1s} \tilde{\gamma}_{1t}^T \tilde{\Sigma}_l \tilde{\gamma}_{1t} / n_l^2\right] + O\left[\sum_{l=1}^2 (\tilde{\gamma}_{1s}^T \tilde{\Sigma}_l \tilde{\gamma}_{1t} / n_l)^2\right]. \quad (\text{B.13})
\end{aligned}$$

Here, under (C-iv), we can evaluate that

$$\begin{aligned}
\sum_{s,t=1}^{q_1} \sum_{u=1}^{q_l} (\tilde{\gamma}_{1s}^T \tilde{\gamma}_{lu})^2 \tilde{\gamma}_{1t}^T \tilde{\gamma}_{lu} \tilde{\gamma}_{1t}^T \tilde{\mu} M_{lu} / n_l^2 &= \sum_{u=1}^{q_l} \tilde{\gamma}_{lu}^T \tilde{\Sigma}_1 \tilde{\gamma}_{lu} \tilde{\gamma}_{lu}^T \tilde{\Sigma}_1 \tilde{\mu} M_{lu} / n_l^2 \\
&= O\left[\|\tilde{\mu}^T \tilde{\Sigma}_1^{1/2}\| \sum_{u=1}^{q_l} \|\tilde{\gamma}_{lu}^T \tilde{\Sigma}_1^{1/2}\| \tilde{\gamma}_{lu}^T \tilde{\Sigma}_1 \tilde{\gamma}_{lu} / n_l^2\right] \\
&= O\left[\|\tilde{\mu}^T \tilde{\Sigma}_1^{1/2}\| \text{tr}(\tilde{\Sigma}_1 \tilde{\Sigma}_l)^{1/2} \left\{ \sum_{u=1}^{q_l} (\tilde{\gamma}_{lu}^T \tilde{\Sigma}_1 \tilde{\gamma}_{lu})^2 \right\}^{1/2} / n_l^2\right] \\
&= O[\{\tilde{\mu}^T \tilde{\Sigma}_1 \tilde{\mu} + \text{tr}(\tilde{\Sigma}_1 \tilde{\Sigma}_l)\} \text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_l)^2\}^{1/2} / n_l^2] = o(\delta_1^4), \quad l = 1, 2 \quad (\text{B.14})
\end{aligned}$$

from the fact that $\sum_{u=1}^{q_l} (\tilde{\gamma}_{lu}^T \tilde{\Sigma}_1 \tilde{\gamma}_{lu})^2 \leq \sum_{u,w=1}^{q_l} (\tilde{\gamma}_{lu}^T \tilde{\Sigma}_1 \tilde{\gamma}_{lw})^2 = \text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_l)^2\} = o(n_l^2 \delta_1^4)$ under (C-iv). Then, by combining (B.12) and (B.13) with (B.14), under (A-ii), (C-iv) and

(C-vii), for any $\tau > 0$, we have that as $m \rightarrow \infty$

$$\sum_{s=1}^{q_1} \frac{E(w_s^4)}{\tau} = O\left[\frac{\sum_{l=1}^2 \text{tr}\{(\tilde{\Sigma}_1 \tilde{\Sigma}_l)^2\}/n_l^2 + \sum_{s=1}^{q_1} (\tilde{\gamma}_{1s}^T \tilde{\mu})^4}{\delta_1^4}\right] \rightarrow 0 \quad \text{and}$$

$$P\left(\left|\sum_{s=1}^{q_1} w_s^2 - 1\right| \geq \tau\right) \leq \frac{\sum_{s,t=1}^{q_1} E(w_s^2 w_t^2) - 1}{\tau^2} = O\left[\sum_{s=1}^{q_1} E(w_s^4)\right] + o(1) \rightarrow 0,$$

so that $\sum_{s=1}^{q_1} E\{w_s^2 I(w_s^2 \geq \tau)\} \leq \sum_{s=1}^{q_1} E(w_s^4)/\tau \rightarrow 0$ and $\sum_{s=1}^{q_1} w_s^2 = 1 + o_P(1)$. Hence, by using the martingale central limit theorem, we obtain that $\sum_{s=1}^{q_1} w_s \Rightarrow N(0, 1)$ as $m \rightarrow \infty$ under (A-ii), (C-iv) and (C-vii). We conclude the result when $i = 1$. For the case when $i = 2$, we can have the same arguments. The proof is completed. \square

Proofs of Corollaries 3.1 and 3.2. From Theorems 3.1 and 3.2 and Proposition 3.1, we can claim Corollaries 3.1 and 3.2 straightforwardly. \square

Lemma B.3. Assume that when $\mathbf{x}_0 \in \pi_i$ for $i = 1, 2$

$$\text{tr}\{(\mathbf{x}_0 - \boldsymbol{\mu}_i)(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T - \boldsymbol{\Sigma}_i\}(\hat{\mathbf{B}}_1 - \hat{\mathbf{B}}_2) = o_P(\kappa); \quad (\text{B.15})$$

$$\text{tr}\{\boldsymbol{\Sigma}_i(\hat{\mathbf{B}}_1 - \hat{\mathbf{B}}_2)\} - \log|\hat{\mathbf{A}}_1 \mathbf{A}_1^{-1}| + \log|\hat{\mathbf{A}}_2 \mathbf{A}_2^{-1}| = o_P(\kappa); \quad \text{and} \quad (\text{B.16})$$

$$\{2(\mathbf{x}_0 - \boldsymbol{\mu}_i) + (-1)^{i+1} \boldsymbol{\mu}_{12}\}^T \hat{\mathbf{B}}_{i'} \boldsymbol{\mu}_{12} = o_P(\kappa) \quad (i' \neq i) \quad (\text{B.17})$$

$$\text{and } (p/n_l^{1/2}) \|\hat{\mathbf{B}}_l\| = o_P(\kappa), \quad l = 1, 2,$$

where $\kappa = \Delta_{\min}$ or $\kappa = \delta_{\min}$. Then, (4.1) holds.

Proof of Lemma B.3. We consider the case when $\mathbf{x}_0 \in \pi_1$. We have that

$$\begin{aligned} & W_1(\hat{\mathbf{A}}_1) - W_1(\mathbf{A}_1) - W_2(\hat{\mathbf{A}}_2) + W_2(\mathbf{A}_2) \\ &= \text{tr}\{(\mathbf{x}_0 - \boldsymbol{\mu}_1)(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T - \boldsymbol{\Sigma}_1\}(\hat{\mathbf{B}}_1 - \hat{\mathbf{B}}_2) \\ &\quad + \text{tr}\{\boldsymbol{\Sigma}_1(\hat{\mathbf{B}}_1 - \hat{\mathbf{B}}_2)\} - \log|\hat{\mathbf{A}}_1 \mathbf{A}_1^{-1}| + \log|\hat{\mathbf{A}}_2 \mathbf{A}_2^{-1}| \\ &\quad + \sum_{l=1}^2 (-1)^{l+1} \text{tr}\{[2(\mathbf{x}_0 - \boldsymbol{\mu}_1 - (\bar{\mathbf{x}}_{ln_l} - \boldsymbol{\mu}_1)/2)(\boldsymbol{\mu}_1 - \bar{\mathbf{x}}_{ln_l})^T - \mathbf{S}_{ln_l}/n_l]\hat{\mathbf{B}}_l\}. \end{aligned}$$

Note that $\text{tr}(\mathbf{S}_{ln_l}) = O_P(p)$, $\|\bar{\mathbf{x}}_{ln_l} - \boldsymbol{\mu}_1\|^2 \leq \|\bar{\mathbf{x}}_{ln_l} - \boldsymbol{\mu}_l\|^2 + \|\boldsymbol{\mu}_l - \boldsymbol{\mu}_1\|^2 = \|\boldsymbol{\mu}_l - \boldsymbol{\mu}_1\|^2 + O_P(p/n_l)$ and $\|\mathbf{x}_0 - \boldsymbol{\mu}_1 - (\bar{\mathbf{x}}_{ln_l} - \boldsymbol{\mu}_1)/2\|^2 \leq \|\mathbf{x}_0 - \boldsymbol{\mu}_1\|^2 + \|\bar{\mathbf{x}}_{ln_l} - \boldsymbol{\mu}_l\|^2 + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_l\|^2 = O_P(p)$, $l = 1, 2$, from the facts that $E(\|\mathbf{x}_0 - \boldsymbol{\mu}_1\|^2) = \text{tr}(\boldsymbol{\Sigma}_1)$, $E\{\text{tr}(\mathbf{S}_{ln_l})\} = \text{tr}(\boldsymbol{\Sigma}_l)$, $E(\|\bar{\mathbf{x}}_{ln_l} - \boldsymbol{\mu}_l\|^2) = \text{tr}(\boldsymbol{\Sigma}_l)/n_l$, $\text{tr}(\boldsymbol{\Sigma}_i) = O(p)$, $i = 1, 2$, and $\|\boldsymbol{\mu}_{12}\|^2 = O(p)$. Then, we have that for $l = 1, 2$

$$\begin{aligned} & |\text{tr}\{[2(\mathbf{x}_0 - \boldsymbol{\mu}_1 - (\bar{\mathbf{x}}_{ln_l} - \boldsymbol{\mu}_1)/2)(\boldsymbol{\mu}_l - \bar{\mathbf{x}}_{ln_l})^T - \mathbf{S}_{ln_l}/n_l]\hat{\mathbf{B}}_l]| \\ & \leq 2\|\mathbf{x}_0 - \boldsymbol{\mu}_1 - (\bar{\mathbf{x}}_{ln_l} - \boldsymbol{\mu}_1)/2\| \cdot \|\bar{\mathbf{x}}_{ln_l} - \boldsymbol{\mu}_l\| \cdot \|\hat{\mathbf{B}}_l\| + \text{tr}(\mathbf{S}_{ln_l})\|\hat{\mathbf{B}}_l\|/n_l \\ & = O_P\{(p/n_l^{1/2})\|\hat{\mathbf{B}}_l\|\}. \end{aligned}$$

Also, we have that $|(\bar{\mathbf{x}}_{2n_2} - \boldsymbol{\mu}_2)^T \hat{\mathbf{B}}_2 \boldsymbol{\mu}_{12}| = O_P\{(p/n_2^{1/2})\|\hat{\mathbf{B}}_2\|\}$. Thus it holds that

$$\begin{aligned} & \sum_{l=1}^2 (-1)^{l+1} \text{tr}[\{2(\mathbf{x}_0 - \boldsymbol{\mu}_1 - (\bar{\mathbf{x}}_{ln_l} - \boldsymbol{\mu}_1)/2)(\boldsymbol{\mu}_1 - \bar{\mathbf{x}}_{ln_l})^T - \mathbf{S}_{ln_l}/n_l\} \hat{\mathbf{B}}_l] \\ &= -\{2(\mathbf{x}_0 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_{12}\}^T \hat{\mathbf{B}}_2 \boldsymbol{\mu}_{12} + O_P\{(p/n_1^{1/2})\|\hat{\mathbf{B}}_1\| + (p/n_2^{1/2})\|\hat{\mathbf{B}}_2\|\}. \end{aligned}$$

Hence, it concludes the result when $\mathbf{x}_0 \in \pi_1$. For the case when $\mathbf{x}_0 \in \pi_2$, we can have the same arguments. The proof is completed. \square

Proofs of Propositions 4.1 and 4.2. We consider the case when $\mathbf{x}_0 \in \pi_1$. Similar to Proof of Lemma B.3, we can claim that $|\{2(\mathbf{x}_0 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_{12}\}^T \hat{\mathbf{B}}_2 \boldsymbol{\mu}_{12}| \leq \|2(\mathbf{x}_0 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_{12}\| \cdot \|\boldsymbol{\mu}_{12}\| \cdot \|\hat{\mathbf{B}}_2\| = O_P(p^{1/2} \|\boldsymbol{\mu}_{12}\| \cdot \|\hat{\mathbf{B}}_2\|) = O_P(p \|\hat{\mathbf{B}}_2\|)$ because $\|\boldsymbol{\mu}_{12}\|^2 = O(p)$ and $\|2(\mathbf{x}_0 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_{12}\|^2 = O_P(p)$. Thus, (B.17) holds under (C-viii) or (C-ix). Note that (B.15) and (B.16) naturally hold when $\hat{\mathbf{A}}_1 = \hat{\mathbf{A}}_2$ and $\mathbf{A}_1 = \mathbf{A}_2$. Hence, from Lemma B.3, it concludes the result of Proposition 4.2 when $\mathbf{x}_0 \in \pi_1$.

Next, we consider (B.15) and the first term of (B.16). We have that for $l = 1, 2$

$$\begin{aligned} |\text{tr}(\boldsymbol{\Sigma}_1 \hat{\mathbf{B}}_l)| &\leq \text{tr}(\boldsymbol{\Sigma}_1) \|\hat{\mathbf{B}}_l\| = O_P(p \|\hat{\mathbf{B}}_l\|) \text{ and} \\ |\text{tr}\{(\mathbf{x}_0 - \boldsymbol{\mu}_1)(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T - \boldsymbol{\Sigma}_1\} \hat{\mathbf{B}}_l| &\leq \|\mathbf{x}_0 - \boldsymbol{\mu}_1\|^2 \|\hat{\mathbf{B}}_l\| + \text{tr}(\boldsymbol{\Sigma}_1) \|\hat{\mathbf{B}}_l\| = O_P(p \|\hat{\mathbf{B}}_l\|). \end{aligned}$$

Finally, we consider $\log |\hat{\mathbf{A}}_l \mathbf{A}_l^{-1}|$, $l = 1, 2$, in (B.16). Let \mathbf{e}_p be an arbitrary (random) p -vector such that $\|\mathbf{e}_p\| = 1$. Note that $\|\mathbf{e}_p^T \mathbf{A}_l^{-1/2}\| \in (0, \infty)$ as $p \rightarrow \infty$ under $\lambda(\mathbf{A}_l) \in (0, \infty)$ as $p \rightarrow \infty$. Thus we have that

$$\mathbf{e}_p^T \mathbf{A}_l^{-1/2} \hat{\mathbf{B}}_l \mathbf{A}_l^{-1/2} \mathbf{e}_p = \mathbf{e}_p^T \mathbf{A}_l^{-1/2} \hat{\mathbf{A}}_l \mathbf{A}_l^{-1/2} \mathbf{e}_p - 1 = O_P(\|\hat{\mathbf{B}}_l\|),$$

so that $\lambda_{\min}(\mathbf{A}_l^{-1/2} \hat{\mathbf{A}}_l \mathbf{A}_l^{-1/2}) - 1 = O_P(\|\hat{\mathbf{B}}_l\|)$ and $\lambda_{\max}(\mathbf{A}_l^{-1/2} \hat{\mathbf{A}}_l \mathbf{A}_l^{-1/2}) - 1 = O_P(\|\hat{\mathbf{B}}_l\|)$. Hence, under $\|\hat{\mathbf{B}}_l\| = o_P(1)$, it holds that for $l = 1, 2$

$$\log |\hat{\mathbf{A}}_l \mathbf{A}_l^{-1}| = \log |\mathbf{A}_l^{-1/2} \hat{\mathbf{A}}_l \mathbf{A}_l^{-1/2}| = O_P(p \|\hat{\mathbf{B}}_l\|).$$

Note that $\Delta_{\min} = O(p)$ and $\delta_{\min} = O(p)$ under $\lambda(\mathbf{A}_i) \in (0, \infty)$ as $p \rightarrow \infty$ for $i = 1, 2$. Then, under (C-viii), it holds that $\|\hat{\mathbf{B}}_l\| = o_P(1)$ for $l = 1, 2$. Hence, (C-viii) implies (B.15) and (B.16). It concludes the result of Proposition 4.1 when $\mathbf{x}_0 \in \pi_1$. For the case when $\mathbf{x}_0 \in \pi_2$, we can have the same arguments. The proof is completed. \square

Proof of Corollary 4.1. Under (A-i) we have that $\text{Var}\{\text{tr}(\mathbf{S}_{in_i})\} = O(\text{tr}(\boldsymbol{\Sigma}_i^2)/n_i)$, $i = 1, 2$, so that $\text{tr}(\mathbf{S}_{in_i}) = \text{tr}(\boldsymbol{\Sigma}_i) + O_P\{(\text{tr}(\boldsymbol{\Sigma}_i^2)/n_i)^{1/2}\}$. Then, under (C-i') it holds that $\text{tr}(\mathbf{S}_{in_i}) = \text{tr}(\boldsymbol{\Sigma}_i) + o_P(\Delta_{\min}(II)) = \text{tr}(\boldsymbol{\Sigma}_i)\{1 + o_P(1)\}$ and $\text{tr}(\boldsymbol{\Sigma}_i^2)/(n_i p^2) = o(\Delta_{\min}^2(II)/p^2) = o(1)$ for $i = 1, 2$ because $\Delta_{\min}(II) = O(p)$. Thus, we have that under (A-i) and (C-i')

$$\begin{aligned} \|\hat{\mathbf{B}}_i\| &= \|\{p/\text{tr}(\mathbf{S}_{in_i}) - p/\text{tr}(\boldsymbol{\Sigma}_i)\} \mathbf{I}_p\| = \frac{p|\text{tr}(\mathbf{S}_{in_i}) - \text{tr}(\boldsymbol{\Sigma}_i)|}{\text{tr}(\mathbf{S}_{in_i})\text{tr}(\boldsymbol{\Sigma}_i)} \\ &= O_P\{(\text{tr}(\boldsymbol{\Sigma}_i^2)/n_i)^{1/2}/\text{tr}(\mathbf{S}_{in_i})\} = o_P\{\Delta_{\min}(II)/p\} = o_P(1), \end{aligned} \quad (\text{B.18})$$

so that $p \|\hat{\mathbf{B}}_i\| = o_P(\Delta_{\min}(II))$. Note that $\lambda_{\max}(\mathbf{A}_i) = \lambda_{\min}(\mathbf{A}_i) = \text{tr}(\boldsymbol{\Sigma}_i)/p \in (0, \infty)$ as $p \rightarrow \infty$. Thus, from Corollary 2.1 and Proposition 4.1, it concludes the result. \square

Proof of Corollary 4.2. We consider the case when $\mathbf{x}_0 \in \pi_i$. Note that $\text{tr}(\mathbf{S}_{l_{n_l}})/\text{tr}(\mathbf{\Sigma}_l) = 1 + O_P\{(\text{tr}(\mathbf{\Sigma}_l^2)/n_l)^{1/2}/p\} = 1 + o_P(1)$, $l = 1, 2$, and $\text{tr}\{(\mathbf{x}_0 - \boldsymbol{\mu}_i)(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T - \mathbf{\Sigma}_i\} = O_P(\text{tr}(\mathbf{\Sigma}_i^2)^{1/2})$ under (A-i). Also, note that $\text{tr}(\mathbf{\Sigma}_i^2)\text{tr}(\mathbf{\Sigma}_l^2) \leq \lambda_{i1}\lambda_{il}\text{tr}(\mathbf{\Sigma}_i)\text{tr}(\mathbf{\Sigma}_l) = o(n_{\min}\delta_{\min(II)}^2 p^2)$, $l = 1, 2$ under (C-iv'). Then, from (B.18), it holds that for $l = 1, 2$

$$\begin{aligned} & \text{tr}[(\mathbf{x}_0 - \boldsymbol{\mu}_i)(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T - \mathbf{\Sigma}_i]\hat{\mathbf{B}}_l \\ &= p \frac{\text{tr}(\mathbf{\Sigma}_l) - \text{tr}(\mathbf{S}_{l_{n_l}})}{\text{tr}(\mathbf{\Sigma}_l)\text{tr}(\mathbf{S}_{l_{n_l}})} \text{tr}\{(\mathbf{x}_0 - \boldsymbol{\mu}_i)(\mathbf{x}_0 - \boldsymbol{\mu}_i)^T - \mathbf{\Sigma}_i\} \\ &= O_P\{(\text{tr}(\mathbf{\Sigma}_i^2)\text{tr}(\mathbf{\Sigma}_l^2)/n_l)^{1/2}/p\} = o_P(\delta_{\min(II)}), \quad \text{and} \\ & p\|\hat{\mathbf{B}}_l\|/n_l^{1/2} = O_P\{\text{tr}(\mathbf{\Sigma}_l^2)^{1/2}/n_l\} = o_P(\delta_{\min(II)}) \end{aligned} \quad (\text{B.19})$$

under (A-i) and (C-iv'). Similarly, from (B.18), under (A-i) and (C-iv'), we have that for $i' \neq i$

$$\begin{aligned} & \{2(\mathbf{x}_0 - \boldsymbol{\mu}_i) + (-1)^{i+1}\boldsymbol{\mu}_{12}\}^T \hat{\mathbf{B}}_{i'} \boldsymbol{\mu}_{12} \\ &= O_P\{(\boldsymbol{\mu}_{12}^T \mathbf{\Sigma}_i \boldsymbol{\mu}_{12}/n_{i'})^{1/2}\} + O_P\{(\text{tr}(\mathbf{\Sigma}_{i'}^2)/n_{i'})^{1/2}\|\boldsymbol{\mu}_{12}\|^2/p\} \\ &= O_P\{(\lambda_{i1}\|\boldsymbol{\mu}_{12}\|^2/n_{i'})^{1/2}\} + O_P\{(\lambda_{i'1}\|\boldsymbol{\mu}_{12}\|^2/n_{i'})^{1/2}\} = o_P(\delta_{\min(II)}) \end{aligned}$$

from the facts that $\boldsymbol{\mu}_{12}^T \mathbf{\Sigma}_i \boldsymbol{\mu}_{12} \leq \lambda_{i1}\|\boldsymbol{\mu}_{12}\|^2$, $\text{tr}(\mathbf{\Sigma}_{i'}^2) = O(\lambda_{i'1}p)$ and $\|\boldsymbol{\mu}_{12}\|^2 = O(p)$. On the other hand, under (A-i) and (C-iv'), from (B.18), we have that for $l = 1, 2$

$$\begin{aligned} \log\{\text{tr}(\mathbf{\Sigma}_l)/\text{tr}(\mathbf{S}_{l_{n_l}})\} &= (\text{tr}(\mathbf{\Sigma}_l)/\text{tr}(\mathbf{S}_{l_{n_l}}) - 1) + O_P\{(\text{tr}(\mathbf{\Sigma}_l)/\text{tr}(\mathbf{S}_{l_{n_l}}) - 1)^2\} \\ &= (\text{tr}(\mathbf{\Sigma}_l)/\text{tr}(\mathbf{S}_{l_{n_l}}) - 1) + O_P\{\text{tr}(\mathbf{\Sigma}_l^2)/(n_l p^2)\} \\ &= (\text{tr}(\mathbf{\Sigma}_l)/\text{tr}(\mathbf{S}_{l_{n_l}}) - 1) + o_P(\delta_{\min(II)}/p) \end{aligned}$$

from the facts that $\text{tr}(\mathbf{\Sigma}_l^2)/p = O(\lambda_{l1})$ and $\text{tr}(\mathbf{\Sigma}_l)/\text{tr}(\mathbf{S}_{l_{n_l}}) = 1 + o_P(1)$. Then, under (A-i) and (C-iv'), it holds that

$$\text{tr}(\mathbf{\Sigma}_i \hat{\mathbf{B}}_i) - \log|\hat{\mathbf{A}}_i \mathbf{A}_i^{-1}| = p(\text{tr}(\mathbf{\Sigma}_i)/\text{tr}(\mathbf{S}_{i_{n_i}}) - 1) - p \log\{\text{tr}(\mathbf{\Sigma}_i)/\text{tr}(\mathbf{S}_{i_{n_i}})\} = o_P(\delta_{\min(II)}).$$

Similarly, under (A-i) and (C-iv'), we have that

$$\begin{aligned} \text{tr}(\mathbf{\Sigma}_i \hat{\mathbf{B}}_{i'}) - \log|\hat{\mathbf{A}}_{i'} \mathbf{A}_{i'}^{-1}| &= p(\text{tr}(\mathbf{\Sigma}_i)/\text{tr}(\mathbf{\Sigma}_{i'}) - 1)(\text{tr}(\mathbf{\Sigma}_{i'})/\text{tr}(\mathbf{S}_{i'_{n_{i'}}}) - 1) + o_P(\delta_{\min(II)}) \\ &= O_P(|\text{tr}(\mathbf{\Sigma}_i)/\text{tr}(\mathbf{\Sigma}_{i'}) - 1|(\text{tr}(\mathbf{\Sigma}_{i'}^2)/n_{i'})^{1/2}) + o_P(\delta_{\min(II)}). \end{aligned} \quad (\text{B.20})$$

By combining (B.19) to (B.20) with Lemma B.3 and Corollary 3.1, we can claim the result. \square

Proof of Corollary 4.3. We can write that

$$s_{in_i(j)} = n_i s_{oin_i(j)}/(n_i - 1) - n_i(\bar{x}_{ijn_i} - \mu_{ij})^2/(n_i - 1), \quad (\text{B.21})$$

where $s_{oin_i(j)} = \sum_{k=1}^{n_i} (x_{ijk} - \mu_{ij})^2/n_i$. Note that $\limsup_{p \rightarrow \infty} E\{\exp(t_{ij}|(x_{ijk} - \mu_{ij})^2 - \sigma_{i(j)}|/\eta_{i(j)}^{1/2})\} \leq \limsup_{p \rightarrow \infty} [E\{\exp(t_{ij}|x_{ijk} - \mu_{ij}|^2/\eta_{i(j)}^{1/2})\} + \exp(t_{ij}\sigma_{i(j)}/\eta_{i(j)}^{1/2})] < \infty$ under

(A-iii). Then, under (A-iii), for any x satisfying $x \rightarrow \infty$ and $x = o(n_i^{1/2})$ as $n_i \rightarrow \infty$, we have that as $n_i \rightarrow \infty$

$$P(n_i^{1/2}|s_{oin_i(j)} - \sigma_{i(j)}|/\eta_{i(j)}^{1/2} \geq x) = \exp\left(-\frac{x^2}{2}\{1 + o(1)\}\right).$$

Refer to Chapter 6 in de la Peña, Lai and Shao (2009) for the details of this result. Let $\tau_{1j} = M(\eta_{i(j)} n_i^{-1} \log p)^{1/2}$ for $j = 1, \dots, p$, where $M > 2^{1/2}$. Then, under $n_i^{-1} \log p = o(1)$, it holds that as $p \rightarrow \infty$

$$\begin{aligned} \sum_{j=1}^p P(|s_{oin_i(j)} - \sigma_{i(j)}| \geq \tau_{1j}) &= \sum_{j=1}^p P(n_i^{1/2}|s_{oin_i(j)} - \sigma_{i(j)}|/\eta_{i(j)}^{1/2} \geq M(\log p)^{1/2}) \\ &= \sum_{j=1}^p \exp\left(-\frac{M^2 \log p}{2}\{1 + o(1)\}\right) \rightarrow 0. \end{aligned} \quad (\text{B.22})$$

Next, we consider the second term of (B.21). Let $u_{ij} = t_{ij}(\sigma_{i(j)}/\eta_{i(j)})^{1/2}$ for $j = 1, \dots, p$. Then, we have that for $j = 1, \dots, p$

$$\begin{aligned} &E\{\exp(u_{ij}|x_{oijk}|/\sigma_{i(j)}^{1/2})\} \\ &= E\{\exp(u_{ij}|x_{oijk}|/\sigma_{i(j)}^{1/2})I(|x_{oijk}| \leq 1)\} + E\{\exp(u_{ij}|x_{oijk}|/\sigma_{i(j)}^{1/2})I(|x_{oijk}| > 1)\} \\ &\leq \exp(u_{ij}/\sigma_{i(j)}^{1/2}) + E\{\exp(u_{ij}x_{oijk}^2/\sigma_{i(j)}^{1/2})\} \leq \exp(u_{ij}/\sigma_{i(j)}^{1/2}) + E\{\exp(t_{is}x_{oijk}^2/\eta_{i(j)}^{1/2})\}, \end{aligned}$$

so that $\limsup_{p \rightarrow \infty} E\{\exp(u_{ij}|x_{oijk}|/\sigma_{i(j)}^{1/2})\} < \infty$ under (A-iii). Thus, in a way similar to (B.22), we have that

$$\sum_{j=1}^p P(|\bar{x}_{ijn_i} - \mu_{ij}| \geq \tau_{2j}) = \sum_{j=1}^p P(n_i^{1/2}|\bar{x}_{ijn_i} - \mu_{ij}|/\sigma_{i(j)}^{1/2} \geq M(\log p)^{1/2}) \rightarrow 0 \quad (\text{B.23})$$

for $\tau_{2j} = M(\sigma_{i(j)} n_i^{-1} \log p)^{1/2}$, $j = 1, \dots, p$. By combining (B.22) and (B.23) with (B.21), under $n_i^{-1} \log p = o(1)$ and (A-iii), we have that

$$\begin{aligned} &\sum_{j=1}^p P\{|s_{in_i(j)} - n_i \sigma_{i(j)}|/(n_i - 1) \geq n_i(\tau_{1j} + \tau_{2j}^2)/(n_i - 1)\} \\ &\leq \sum_{j=1}^p P(|s_{oin_i(j)} - \sigma_{i(j)}| + |\bar{x}_{ijn_i} - \mu_{ij}|^2 \geq \tau_{1j} + \tau_{2j}^2) \\ &\leq \sum_{j=1}^p P(|s_{oin_i(j)} - \sigma_{i(j)}| \geq \tau_{1j}) + \sum_{j=1}^p P(|\bar{x}_{ijn_i} - \mu_{ij}|^2 \geq \tau_{2j}^2) \rightarrow 0. \end{aligned}$$

Note that $n_i \sigma_{i(j)}/(n_i - 1) = \sigma_{i(j)} + o(n_i^{-1/2})$ and $\tau_{2j}^2 = o(\tau_{1j})$ under $n_i^{-1} \log p = o(1)$. Thus we have that $\max_{j=1, \dots, p} \{|s_{in_i(j)} - \sigma_{i(j)}|\} = O_P(\max_{j=1, \dots, p} \tau_{1j})$ under $n_i^{-1} \log p = o(1)$ and (A-iii), so that

$$\max_{j=1, \dots, p} \{|s_{in_i(j)} - \sigma_{i(j)}|\} = O_P\{(n_i^{-1} \log p)^{1/2}\}. \quad (\text{B.24})$$

Then, for $i = 1, 2$, it holds that under $n_i^{-1} \log p = o(1)$

$$\begin{aligned} \|\hat{\mathbf{B}}_i\| &= \|\mathbf{S}_{i(d)}^{-1} - \mathbf{\Sigma}_{i(d)}^{-1}\| = \max_{j=1, \dots, p} \{|s_{in_i(j)} - \sigma_{i(j)}| / (s_{in_i(j)} \sigma_{i(j)})\} \\ &= O_P\{(n_i^{-1} \log p)^{1/2}\} = o_P(1). \end{aligned} \quad (\text{B.25})$$

Then, it follows that (C-i') holds under (4.4). From the facts that $\Delta_{\min(III)} = O(p)$, note that $n_{\min}^{-1} \log p = o(1)$ under (4.4). Then, by combining (B.25) with Proposition 4.1 and Corollary 2.1, we can claim the result of Corollary 4.3. \square

Proofs of Corollary 4.4. First, note that $s_{n(j)} - \sigma_{(j)} = \sum_{i=1}^2 (n_i - 1)(s_{in_i(j)} - \sigma_{i(j)}) / (\sum_{i=1}^2 n_i - 2)$. From (B.24), we can claim that $\max_{j=1, \dots, p} \{|s_{n(j)} - \sigma_{(j)}|\} = O_P\{(n_{\min}^{-1} \log p)^{1/2}\}$ under $n_{\min}^{-1} \log p = o(1)$ and (A-iii). Thus it follows that $\|\mathbf{S}_{n(d)}^{-1} - \mathbf{\Sigma}_{(d)}^{-1}\| = O_P\{(n_{\min}^{-1} \log p)^{1/2}\}$. Note that $\Delta_{(III)} / \|\boldsymbol{\mu}_{12}\|^2 \in (0, \infty)$ as $p \rightarrow \infty$. Then, by combining Theorem 2.1 with Propositions 2.1 and 4.2, we can claim the result of Corollary 4.4. \square

Proofs of Corollary 4.5. Let $\mathbf{S}_{oin_i} = \sum_{k=1}^{n_i} (\mathbf{x}_{ik} - \boldsymbol{\mu}_i)(\mathbf{x}_{ik} - \boldsymbol{\mu}_i)^T / n_i$ and denote its (r, s) element by $s_{oin_i(rs)}$ for $r, s = 1, \dots, p$. Let $u_{i(rs)} = \min\{t_{ir}/\eta_{i(r)}^{1/2}, t_{is}/\eta_{i(s)}^{1/2}\} \eta_{i(rs)}^{1/2}$ for $r, s = 1, \dots, p$. Then, we have that for $r, s = 1, \dots, p$

$$\begin{aligned} &E\{\exp(u_{i(rs)} |x_{oirk} x_{oisk} - \sigma_{i(rs)}| / \eta_{i(rs)}^{1/2})\} \\ &\leq E[\exp\{u_{i(rs)}(x_{oirk}^2/2 + x_{oisk}^2/2 + \sigma_{i(rs)}) / \eta_{i(rs)}^{1/2}\}] \\ &\leq \exp(u_{i(rs)} \sigma_{i(rs)} / \eta_{i(rs)}^{1/2}) E[\exp\{t_{ir} x_{oirk}^2 / (2\eta_{i(r)}^{1/2})\} \exp\{t_{is} x_{oisk}^2 / (2\eta_{i(s)}^{1/2})\}] \\ &\leq \exp(u_{i(rs)} \sigma_{i(rs)} / \eta_{i(rs)}^{1/2}) [E\{\exp(t_{ir} x_{oirk}^2 / \eta_{i(r)}^{1/2})\} E\{\exp(t_{is} x_{oisk}^2 / \eta_{i(s)}^{1/2})\}]^{1/2}, \end{aligned}$$

so that $\limsup_{p \rightarrow \infty} E\{\exp(u_{i(rs)} |x_{oirk} x_{oisk} - \sigma_{i(rs)}| / \eta_{i(rs)}^{1/2})\} < \infty$ under (A-iii). Note that $s_{in_i(rs)} = n_i s_{oin_i(rs)} / (n_i - 1) - n_i(\bar{x}_{irn_i} - \mu_{ir})(\bar{x}_{isn_i} - \mu_{is}) / (n_i - 1)$, where $s_{in_i(rs)}$ is the (r, s) element of \mathbf{S}_{in_i} . Also, note that $\eta_{i(rs)} \in (0, \infty)$ as $p \rightarrow \infty$ under (A-iii) and $\liminf_{p \rightarrow \infty} \eta_{i(rs)} > 0$ for all r, s , from the fact that $\eta_{i(rs)} \leq \{(\eta_{i(r)} + \sigma_{i(r)}^2)(\eta_{i(s)} + \sigma_{i(s)}^2)\}^{1/2}$. In a way similar to (B.22) and (B.23), under $n_i^{-1} \log p = o(1)$, (A-iii) and $\liminf_{p \rightarrow \infty} \eta_{i(rs)} > 0$ for all r, s , we have that

$$\begin{aligned} &\sum_{r,s=1}^p P\{|s_{in_i(rs)} - n_i \sigma_{i(rs)}| / (n_i - 1) \geq n_i(\tau_{1(rs)} + \tau_{2(rs)}) / (n_i - 1)\} \\ &\leq \sum_{r,s=1}^p \{P(|s_{oin_i(rs)} - \sigma_{i(rs)}| \geq \tau_{1(rs)}) + P(|\bar{x}_{irn_i} - \mu_{ir}| |\bar{x}_{isn_i} - \mu_{is}| \geq \tau_{2(rs)})\} \\ &\leq \sum_{r,s=1}^p P(|\bar{x}_{irn_i} - \mu_{ir}|^2 + |\bar{x}_{isn_i} - \mu_{is}|^2 \geq \tau_{2(rs)}) + o(1) \rightarrow 0 \end{aligned}$$

for $\tau_{1(rs)} = M(\eta_{i(rs)} n_i^{-1} \log p)^{1/2}$ and $\tau_{2(rs)} = M^2\{(\sigma_{i(r)} + \sigma_{i(s)}) n_i^{-1} \log p\}$, $r, s = 1, \dots, p$, where $M > 2$. Thus it holds that $\max_{r,s=1, \dots, p} \{|s_{in_i(rs)} - \sigma_{i(rs)}|\} = O_P(\max_{r,s=1, \dots, p} \tau_{1(rs)})$ because $\tau_{2(rs)} = o(\tau_{1(rs)})$, so that

$$\max_{r,s=1, \dots, p} \{|s_{in_i(rs)} - \sigma_{i(rs)}|\} = O_P\{(n_i^{-1} \log p)^{1/2}\}. \quad (\text{B.26})$$

Here, from the equations (A1) and (A2) in Bickel and Levina (2008a), we have that $\|\mathbf{M}\| \leq \max_{s=1,\dots,p} \sum_{t=1}^p |m_{st}|$ for any symmetric matrix \mathbf{M} , where m_{st} is the (s, t) element of \mathbf{M} . From (B.26), we have that

$$\|\mathbf{S}_{in_i} - \boldsymbol{\Sigma}_i\| = O_P\{p(n_i^{-1} \log p)^{1/2}\} = o_P(1) \quad (\text{B.27})$$

under $n_i^{-1} p^2 \log p = o(1)$, (A-iii) and $\liminf_{p \rightarrow \infty} \eta_{i(rs)} > 0$ for all r, s . Then, under $\lambda(\boldsymbol{\Sigma}_i) \in (0, \infty)$ as $p \rightarrow \infty$, we can claim that $\lambda(\mathbf{S}_{in_i}) \in (0, \infty)$ in probability. Thus it holds that $\|\mathbf{e}_p^T \boldsymbol{\Sigma}_i^{-1}\| \in (0, \infty)$ and $\|\mathbf{e}_p^T \mathbf{S}_{in_i}^{-1}\| \in (0, \infty)$ in probability, where \mathbf{e}_p is an arbitrary (random) p -vector such that $\|\mathbf{e}_p\| = 1$. Then, from (B.27), we have that $\mathbf{e}_p^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{S}_{in_i} - \boldsymbol{\Sigma}_i) \mathbf{S}_{in_i}^{-1} \mathbf{e}_p = \mathbf{e}_p^T (\boldsymbol{\Sigma}_i^{-1} - \mathbf{S}_{in_i}^{-1}) \mathbf{e}_p = O_P\{p(n_i^{-1} \log p)^{1/2}\}$ under $n_i^{-1} p^2 \log p = o(1)$, (A-iii) and $\liminf_{p \rightarrow \infty} \eta_{i(rs)} > 0$ for all r, s , so that $\|\hat{\mathbf{B}}_i\| = O_P\{p(n_i^{-1} \log p)^{1/2}\} = o_P(1)$. Note that (C-i') and (C-ii') hold under the conditions of Corollary 4.5. Also, note that $\text{tr}\{(\mathbf{I}_p - \boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'}^{-1})^2\} = O(p)$ ($i' \neq i$) under $\lambda(\boldsymbol{\Sigma}_i) \in (0, \infty)$ as $p \rightarrow \infty$. By combining Corollary 3.2 with Proposition 4.1, we can claim the result of Corollary 4.5. \square

Proof of Corollary 5.1. By using Theorem 5.1, we can claim the result straightforwardly. \square

Proof of Corollary 5.2. Let us write that for $i = 1, 2$

$$W_i(\boldsymbol{\Sigma}_{i(d)}^{-1})_{FS} = \sum_{j \in \mathbf{D}} \{(x_{0j} - \bar{x}_{ij n_i})^2 / \sigma_{i(j)} - s_{in_i(j)} / (\sigma_{i(j)} n_i) + \log \sigma_{i(j)}\}.$$

Note that $E\{W_{i'}(\boldsymbol{\Sigma}_{i'(d)}^{-1})_{FS}\} - E\{W_i(\boldsymbol{\Sigma}_{i(d)}^{-1})_{FS}\} = \Delta_{i(III)}$ ($i' \neq i$) when $\mathbf{x}_0 \in \pi_i$. Also note that $\liminf_{p \rightarrow \infty} \Delta_{\min(III)} / p_* > 0$ under $\liminf_{p \rightarrow \infty} \theta_j > 0$ for all $j \in \mathbf{D}$. If $\lambda_{\max}(\boldsymbol{\Sigma}_{i*}) = o(p_*)$, (C-i') and (C-ii') hold for $\boldsymbol{\Sigma}_{i*}$, $i = 1, 2$. Here, let us write that $\boldsymbol{\Sigma}_{i(d)*} = \text{diag}(\sigma_{i(j_1)}, \dots, \sigma_{i(j_{p_*})})$ and $\mathbf{S}_{i(d)*} = \text{diag}(s_{in_i(j_1)}, \dots, s_{in_i(j_{p_*})})$ for $i = 1, 2$, where $\mathbf{D} = \{j_1, \dots, j_{p_*}\}$. Then, in a way similar to (B.25), under $n_i^{-1} \log p = o(1)$ and (A-iii), it holds that $\|\mathbf{S}_{i(d)*}^{-1} - \boldsymbol{\Sigma}_{i(d)*}^{-1}\| = O_P\{(n_i^{-1} \log p)^{1/2}\}$. Hence, we have that $p_* \|\mathbf{S}_{j(d)*}^{-1} - \boldsymbol{\Sigma}_{j(d)*}^{-1}\| = o_P(\Delta_{\min(III)})$ under $\liminf_{p \rightarrow \infty} \theta_j > 0$ for all $j \in \mathbf{D}$. By combining Corollary 5.1 with Propositions 2.1 and 4.1, we can claim the result. \square

Acknowledgements

Research of the first author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Exploratory Research, Japan Society for the Promotion of Science (JSPS), under Contract Numbers 15H01678 and 26540010. Research of the second author was partially supported by Grant-in-Aid for Young Scientists (B), JSPS, under Contract Number 26800078.

References

Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequential Anal. (Editor's special invited paper)*, **30**, 356–399.

- Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Ann. Inst. Statist. Math.*, **66**, 983–1010.
- Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R. and Korsmeyer, S.J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, **30**, 41–47.
- Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statist. Sinica*, **6**, 311–329.
- Bickel, P.J. and Levina, E. (2004). Some theory for Fisher’s linear discriminant function, ‘naive Bayes’, and some alternatives when there are many more variables than observations. *Bernoulli*, **10**, 989–1010.
- Bickel, P.J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.*, **36**, 199–227.
- Bickel, P.J. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.*, **36**, 2577–2604.
- Cai, T.T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, **106**, 594–607.
- Cai, T.T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.*, **106**, 1566–1577.
- Chan, Y.-B. and Hall, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika*, **96**, 469–478.
- de la Peña, V.H., Lai, T.L. and Shao, Q.M. (2009). *Self-Normalized Processes*. Berlin: Springer-Verlag.
- Dudoit, S., Fridlyand, J. and Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.*, **97**, 77–87.
- Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.*, **36**, 2605–2637.
- Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *J. R. Statist. Soc. Ser. B*, **74**, 745–771.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
- Hall, P., Marron, J.S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc. Ser. B*, **67**, 427–444.

- Huang, S., Tong, T. and Zhao, H. (2010). Bias-corrected diagonal discriminant rules for high-dimensional classification. *Biometrics*, **66**, 1096–1106.
- Li, Q. and Shao, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statist. Sinica*, in press (doi: 10.5705/ss.2013.150).
- Marron, J.S., Todd, M.J. and Ahn, J. (2007). Distance-weighted discrimination. *J. Amer. Statist. Assoc.*, **102**, 1267–1271.
- McLeish, D.L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probab.*, **2**, 620–628.
- Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.*, **39**, 1241–1265.
- Tan, A.K, Naiman, D.Q., Xu, L., Winslow, R.L. and Geman, D. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, **21**, 3896–3904.
- Vapnic, V.N. (1999). *The Nature of Statistical Learning Theory (second ed.)*. New York: Springer-Verlag.
- Yata, K. and Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. *J. Multivariate Anal.*, **122**, 334–354.